2020

Methods for correcting and analyzing gene families

Akshay Yadav
Iowa State University

Methods for correcting and analyzing gene families

by

Akshay Yadav

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Steven Cannon, Co-major Professor
David Fernández-Baca, Co-major Professor
Karin Dorman
Matthew Hufford
Oliver Eulenstein

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2020

DEDICATION

To my wife, Prajakta, my brother, Digvijay, and to Aai and Baba for their unconditional support and encouragement.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS

I would like to start by thanking my major professors, Dr. Steven Cannon and Dr. David Fernández-Baca their valuable guidance and support throughout the course of this research.

I would also like to thank my committee members, Dr. Karin Dorman, Dr. Matthew Hufford, and Dr. Oliver Eulenstein for their critical assessments and reviews of this work during committee meetings. Regular meetings with Dr. Dorman, during the initial stages, were extremely helpful for completion of this work.

A heartfelt thanks to my friends, Pulkit Kanodia, Gaurav Kandoi, Viraj and Bhakti Muthye, Naihui Zhou, Talon Brown, and Surya and Saranya for the much-needed breaks from work and making my stay at Ames a memorable one.

I would also like to express my gratitude to the former and current program coordinators of the BCB program, Trish Stauble and Carla Harris, for their valuable help in all the administrative matters. A special thank you to Trish for organizing all the BCB dinners and social events and helping me navigate the first year of the course.

A big thank you to my in-laws and extended family members, Mr. Shrirang Patwardhan and Mrs. Hema Patwardhan, Mr. Janardhan Lele, and Mr. Ranjit Shinde for their faith in me and keeping their patience with me, over all these years.

Finally, I would like to thank my parents, Mrs. Tilottama Yadav and Mr. Ashok Yadav, my wife, Prajakta Patwardhan, and my brother, Digvijay Yadav for their love, support, and encouragement to follow my dreams.

ABSTRACT

Gene families are groups of genes that have descended from a common ancestral gene present in the species under study. Current, widely used gene family building algorithms are prone to producing incomplete families (under-clustering) or families containing wrong or non-family sequences (over-clustering). In this work, we present a sequence-pair-classification-based method that, first, inspects given families for under-clustering and then predicts the missing sequences for the families using family-specific alignment score cutoffs. We test this method on a set of curated, gold-standard families from the Yeast Gene Order Browser (YGOB) database, including 20 yeast species. To check if the method can detect and correct incomplete families obtained using existing family building methods, we test this method on under-clustered yeast families produced using the OrthoFinder tool. We demonstrate the utility of the pair-classification method in merging small, fragmented legume families into larger families, built using the OrthoFinder tool, from 14 legumes species belonging to subfamily Papilionoideae of the plant family Leguminosae. We provide recommendations on different types of family-specific alignment score cutoffs that can be used for predicting the missing sequences based on the "purity" of under-clustered families and the chosen precision and recall for prediction. Finally, we provide the containerized version of the pair-classification method that can be applied on any given set of gene families.

In addition to the pair-based classification method, we present a simple hidden Markov model (HMM)-based protocol for merging fragmented families and a phylogeny-based protocol for detecting and splitting over-clustered families. We apply these methods for improving the legume gene families built from 14 legumes species belonging to subfamily Papilionoideae of the plant family Leguminosae, using a custom family building method, that utilizes differences

in the synonymous-sites (K_s) in the gene sequences in order to capture the family clusters defined by the whole-genome duplication that occurred in the most recent common ancestor of the subfamily. We also analyze the improvements in the legume families obtained after the application of merging and splitting procedures by comparing the protein domain compositions of the new families against the original families. We also provide the containerized versions of family merging, splitting and scoring methods along with the new set of improved legume families.

We investigate the occurrence of whole-genome duplication events within the Cercidoideae subfamily of the plant family Leguminosae, using evolutionary, phylogenomic, and synteny analyses together with analysis of chromosome counts, from a diverse set of legume species. Based on diverse evidence, we conclude that one of the slow-evolving lineages within Cercidoideae may be unique among legumes in lacking evidence of an independent whole-genome duplication and can be a useful genomic model for the legumes. We are able to show that the genome duplication observed in the other sister lineage within Cercidoideae is most likely due to allotetraploidy involving hybridization between two progenitor species that existed in the Cercidoideae subfamily.

We present a method for tracking protein domain changes in a selected set of species with known phylogenetic relationships, by defining domains as “features” or “descriptors,” and considering the species (target + outgroup) as instances or data-points in a domain feature matrix. Protein domains can be regarded as sections of protein sequences capable of folding independently and performing specific functions that enable protein sequences to evolve through domain shuffling events like domain insertion, deletion, or duplication. We look for features (domains) that are significantly different between the target species and the outgroup species

using a feature selection technique called Mutual-Information (MI) and non-parametric statistical tests (Fisher's exact test/Wilcoxon rank-sum test). We study the domain changes in two large, distinct groups of plant species: legumes (Fabaceae) and grasses (Poaceae), with respect to selected outgroup species, using four types of domain feature matrices: domain content, domain duplication, domain abundance, and domain versatility. The four types of domain feature matrices attempt to capture different aspects of domain changes through which the protein sequences may evolve - i.e. via gain or loss of domains, increase or decrease in the copy number of domains along the sequences, expansion or contraction of domains, or through changes in the number of adjacent domain partners. We report and study the biological functions of the top selected domains from all four feature matrices. In addition, we perform domain-centric Gene Ontology (dcGO) enrichment analysis on all selected domains from all the feature matrices to study the Gene Ontology terms associated with the significantly changing domains in legumes and grasses. We provide a docker container that can be used to perform this analysis on any user-defined sets of species.

CHAPTER 1. GENERAL INTRODUCTION

Homology and Gene Families

Homology is described as the relationship between gene sequences that have descended from a common ancestral gene, mainly through divergent form of evolution [1]. Homology can be broadly divided into two subtypes: Orthology and Paralogy [2, 3]. Orthology is the relationship between any two homologous sequences that have diverged due to speciation. Consequently, the ancestral sequence for orthologous sequences lies in the common ancestor of the species from which the sequences were obtained. Paralogy is defined as the relationship between homologous sequences that have separated due to duplication of the ancestral gene. Therefore, unlike orthologous sequences, paralogous sequences can exist and evolve side-by-side in the same lineage.

Gene families are clusters of homologous sequences, typically from multiple species, where each cluster contains genes that have descended from a single ancestral gene present in the most recent common ancestor or MRCA [2] of the species under consideration. The gene clusters can contain orthologs that have diverged due to speciation events and paralogs that have diverged due to duplication events (local or large-scale) occurring after the separation of the MRCA of the species. Specifically, paralogs that have originated due to duplication events taking place before the separation of MRCA have to be placed in different clusters [4]. Therefore, gene families are constructed with respect to a phylogenetic range, which can be defined by the MRCA node in the species phylogeny or by outgroup species - species that have diverged before the separation of the MRCA. This means that the phylogenies of individual gene families should agree with species relationships under the MRCA and relationship between the MRCA and the outgroups.

Gene Family Building

Since gene families are clusters of sequences, regular clustering techniques such as single-linkage or Markov clustering (MCL) [5] that use alignment statistics from sequence alignment tools such as BLAST [6] are employed for building gene families. An early gene family resource, the COG database [7], built families from five phylogenetically distant lineages by detecting triangles between orthologous sequences from any three lineages and merging any two triangles with a common side. The ortholog triangles were detected using the reciprocal best hits (RBH) technique [8, 9], where two proteomes are searched against each other and sequence pairs that find each other as best hits are considered as orthologs. The InParanoid method [4] uses a similar RBH-based method for building families by first detecting ortholog pairs between any two lineages and then gathering recent paralogs using the alignment statistics of ortholog pairs. MultiParanoid [10], the multi-lineage implementation of InParanoid, uses single-linkage clustering for building families for multiple lineages from pairwise lineage results. High-throughput family construction methods such as OrthoMCL [11] and OrthoFinder [12] use faster and more effective clustering methods such as MCL that use normalized BLAST E-values or alignment scores to cluster sequences into families. These high-throughput methods generally use common values of clustering parameters for building all the families on a species-wide scale. It may not be appropriate to use common parameter values for building all the families since different gene families evolve at different rates [13–16]. Using a single value of a clustering parameter that controls the granularity of family clusters could be too stringent for families that evolve faster, resulting into the corresponding clusters missing true family sequences (under-clustering), or could be too relaxed for families that evolve slower, resulting into the corresponding clusters containing wrong or non-family sequences (over-clustering).

In **Chapter 2**, we attempt to solve the under-clustering problem by training family-tailored classification models, based on analysis of sequence pairs. The method first inspects a given family for under-clustering, and subsequently attempts to predict missing sequences for the family using family-specific alignment score cutoffs obtained in the training step. The training step consists of repetitive model building and testing where, during each iteration, a combined set of sequences from the given family along with a selected set of closest non-family sequences is randomly split into training and testing parts. The training set of sequence pairs from the family are used to build a hidden Markov model (HMM) [17], which is then tested to recognize the "correct" family sequences from the testing part – which contains both family and non-family sequences. We assess the effectiveness of this method in detecting complete families and correcting artificially modified yeast families. We also show the effectiveness of this method in correcting yeast families and merging small fragmented legume families into larger families, produced using the OrthoFinder tool.

In **Chapter 3**, we present a simple HMM-based [17] protocol for merging fragmented and under-clustered families and a tree-based [18] family scoring and splitting method for correcting over-clustered families. Both the methods leverage the outgroup-based and phylogenetic properties of gene families for merging and splitting. The family merging strategy is based on a two-way HMM-based database search procedure in which missing sequences are predicted for each family using their family HMMs and the outgroup sequences. Subsequently, a simple overlap rule is used to merge families using the predicted missing sequences. The tree-based family scoring and splitting method is based on detecting monophyletic ingroup sequence clades in relation to outgroup sequences. We demonstrate the effectiveness of these methods in improving the legumes families hosted at legumeinfo.org in 2019 [19].

The Legume Family

The legume family (Leguminosae, Fabaceae), comprising over 750 genera and 20,000 species, is the third largest family of flowering plants [20]. The family diverged into six subfamilies (Papilionoideae, Caesalpinioideae, Detarioideae, Cercidoideae, Dialioideae, Duparquetioideae) shortly after its origin ~59 - 64 million years ago (Mya) [21]. The largest subfamily, Papilionoideae, which includes familiar legumes such as soybean (*Glycine max*), mungbean (*Vigna radiata*), pea (*Pisum sativum*), and pigeonpea (*Cajanus cajan*), was affected by an early whole-genome duplication (WGD) prior to its diversification, about 55 million years ago (Mya) [22]. In addition, some genera such as *Glycine* and *Lupinus* have also undergone independent, lineage specific WGDs.

Gene families hosted at legumeinfo.org in 2019 [19] are built from 14 legume proteomes all belonging to the subfamily Papilionoideae. The families were built using a custom family construction method in order to circumscribe the family clusters such that each family captures all legume orthologs and paralogs deriving from speciation events and papilionoid WGD, but not sequences deriving from earlier WGD events. The method used a combination of homology-filtering based on per-species synonymous site changes (K_s), comparisons with outgroup species, Markov clustering, and progressive refinements of family Hidden Markov Models (HMMs).We use the family merging and splitting methods described in **Chapter 3** to improve the K_s-based legume families. We also design and employ a protein-domain-composition-based family scoring method to check the improvements in K_s-based legume families after the application family merging and splitting procedures.

In case of the subfamily Cercidoideae, the status and timing of WGD has been unresolved. A WGD signal was reported by Cannon et al. (2015) [22] in the genus *Bauhinia*, based on synonymous substitution distributions (K_s peaks for duplication and speciation) from

transcriptome sequence - but no conclusive evidence was found for a WGD in the sister genus *Cercis*. In **Chapter 4**, we investigate the WGD and allopolyploidy events within Cercidoideae using the K_s-based families, containing sequences from Cercidoideae, Caesalpinioideae and Papilionoideae. We used a custom method for investigating a large number of tree topologies containing *Cercis* and *Bauhinia* sequences, along with results from analysis of K_s peaks, synteny analysis, species representation within gene families and gene duplication patterns, to show the lack of WGD in *Cercis*, and a WGD in *Bauhinia*, likely resulting from hybridization between the *Cercis* progenitor and a second diploid species within the Cercidoideae.

Protein Domains

Protein domains - sections of protein sequences with the ability to fold and function independently - can be considered as “lego bricks” that can be recombined in various ways to build new proteins [23, 24]. These are independent evolutionary units of proteins that enable proteins to evolve in a modular fashion through domain insertion, deletion, duplication, or substitution, in addition to evolution through point mutations [25, 26].Therefore, tracking gain or loss of particular domains in a group of species can provide means of understanding trait evolution in those species [27, 28]. Similarly, protein domains can duplicate along the sequences and significantly different duplicate counts of certain domains in a target set of species relative to an outgroup set may provide some useful information about the increase or decrease in the functions associated with those domains [29, 30]. Protein domains can also increase or decrease in numbers through duplications or deletions of sequences carrying them. These changes can be useful in inferring biological functions [31]. Protein domains can also be “versatile” in partnering with multiple different domains along the protein sequences and domains that increase or decrease in their versatility can be useful in studying the evolution of associated functions [23,

32, 33]. In **Chapter 5**, we leverage these domain properties to describe whole species, using protein domains [34] as "features". We calculate domain feature matrices with rows representing species, columns representing the protein domains and the cells containing domain feature values for the respective species. We build and analyze different feature matrices to study the changes in four domain properties: content, duplication, abundance and versatility, in two sets of plant species: legumes (Fabaceae) and grasses (Poaceae). We analyze these feature matrices using Mutual-Information (MI) based feature selection and statistical testing techniques to filter out protein domains that have significantly different properties in legumes and grasses as compared to their respective outgroups. MI measures mutual dependence between two random variables by quantifying the amount of information communicated about one random variable from another random variable [35]. MI has been routinely used for selecting meaningful features, in classification and pattern recognition problems [36–38]. Here, we used MI to quantify the mutual dependence between domain feature values and the classification between target and outgroup species. We also employed tests for significance of differences in domain feature values between the target and outgroup species. We applied the Fisher's exact test [39] on feature matrices containing discrete values, and the Wilcoxon rank-sum test [40] on feature matrices containing continuous values. We also report and study the functions of the top significantly different domains and the significantly enriched Gene Ontology (GO) terms found in all the significantly different domains from all four feature matrices, from both the species sets.

References

1. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

2. Fitch WM (2000) Homology: a personal view on some of the problems. Trends Genet 16:227–231

3. Sonnhammer EL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. TRENDS Genet 18:619–620

4. Remm M, Storm CEV, Sonnhammer ELL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052

5. Van Dongen SM (2000) Graph clustering by flow simulation. PhD Thesis

6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

7. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

8. Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. Nature 411:1046

9. Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 12:962–968

10. Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22:e9–e15

11. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res 13:2178–2189

12. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

13. Ohta T (2000) Evolution of gene families. Gene 259:45–52

14. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. PloS One 1:e85

15. De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22:1269–1271

16. Trachana K, Larsson TA, Powell S, Chen W-H, Doerks T, Muller J, Bork P (2011) Orthology prediction methods: a quality assessment using curated protein families. BioEssays News Rev Mol Cell Dev Biol 33:769–780

17. Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6:361–365

18. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5:e9490

19. Stai JS, Yadav A, Sinou C, Bruneau A, Doyle JJ, Fernández-Baca D, Cannon SB (2019) *Cercis*: A Non-polyploid Genomic Relic Within the Generally Polyploid Legume Family. Front Plant Sci. https://doi.org/10.3389/fpls.2019.00345

20. Lewis GP (2005) Legumes of the World. Royal Botanic Gardens Kew

21. Azani N, Babineau M, Bailey CD, et al (2017) A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). TAXON 66:44–77

22. Cannon SB, McKain MR, Harkess A, et al (2015) Multiple Polyploidy Events in the Early Radiation of Nodulating and Nonnodulating Legumes. Mol Biol Evol 32:193–210

23. Vogel C, Teichmann SA, Pereira-Leal J (2005) The Relationship Between Domain Duplication and Recombination. J Mol Biol 346:355–365

24. Das S, Smith TF (2000) Identifying nature's protein Lego set. Adv Protein Chem 54:159–184

25. Liu J, Rost B (2004) CHOP: parsing proteins into structural domains. Nucleic Acids Res 32:W569–W571

26. Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J (2005) The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci CMLS 62:435–445

27. Nasir A, Kim KM, Caetano-Anollés G (2014) Global Patterns of Protein Domain Gain and Loss in Superkingdoms. PLOS Comput Biol 10:e1003452

28. Buljan M, Frankish A, Bateman A (2010) Quantifying the mechanisms of domain gain in animal proteins. Genome Biol 11:R74

29. Björklund ÅK, Ekman D, Elofsson A (2006) Expansion of Protein Domain Repeats. PLOS Comput Biol 2:e114

30. Yasutake Y, Watanabe S, Yao M, Takada Y, Fukunaga N, Tanaka I (2002) Structure of the Monomeric Isocitrate Dehydrogenase: Evidence of a Protein Monomerization by a Domain Duplication. Structure 10:1637–1648

31. Vogel C, Chothia C (2006) Protein Family Expansions and Biological Complexity. PLOS Comput Biol 2:e48

32. Basu MK, Poliakov E, Rogozin IB (2009) Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform 10:205–216

33. Forslund K, Sonnhammer ELL (2012) Evolution of Protein Domain Architectures. In: Anisimova M (ed) Evol. Genomics Stat. Comput. Methods Vol. 2. Humana Press, Totowa, NJ, pp 187–216

34. El-Gebali S, Mistry J, Bateman A, et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432

35. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69:066138

36. Amiri F, Rezaei Yousefi M, Lucas C, Shakery A, Yazdani N (2011) Mutual information-based feature selection for intrusion detection systems. J Netw Comput Appl 34:1184–1199

37. Kraskov A, Stögbauer H, Andrzejak RG, Grassberger P (2003) Hierarchical Clustering Based on Mutual Information. ArXivq-Bio0311039

38. Beraha M, Metelli AM, Papini M, Tirinzoni A, Restelli M (2019) Feature Selection via Mutual Information: New Theoretical Insights. ArXiv190707384 Cs Stat

39. Fisher RA (1922) On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. https://doi.org/10.2307/2340521

40. Mann HB, Whitney DR (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat 18:50–60

CHAPTER 2. METHODS FOR ANALYZING, COMPARING AND CORRECTING GENE FAMILIES

Akshay Yadav, David Fernández-Baca, Steven B. Cannon

Modified from a manuscript to be submitted to a peer reviewed journal

Abstract

Gene families are groups of genes that have descended from a common ancestral gene present in the set of species under study. Current, widely used gene family building algorithms can produce family clusters that may be fragmented or missing true family sequences (under-clustering) or can also produce family clusters with unrelated or "wrong" sequences (over-clustering). In this work, we present (1) a machine-learning classification method that determines per-family homology parameters and detects and corrects under-clustered gene families; (2) a method for comparing family-sets or clustering solutions obtained using different family building methods, and (3) a method for scoring families relative to a known species phylogeny, to detect over-clustering in families. We tested the under-clustering detection and correction method on a set of curated, gold-standard families from the Yeast Gene Order Browser (YGOB) database, including 20 yeast species, as well as a test set of intentionally under-clustered ("deficient") families derived from the YGOB families. We used the family-sets comparison method to compare the yeast families built from OrthoFinder to the reference yeast families from the YGOB database, to detect under-clustered and over-clustered yeast families produced using OrthoFinder. Also, to check if the machine learning method can correct incomplete families obtained using existing family building methods, we applied it to synthetic under-clustered yeast families. We also analyzed 14,663 legume families built using the

OrthoFinder program, with 14 species from the legume plant family. Using the machine learning evaluation method, we were able to identify 1,665 OrthoFinder legume families that were missing one or more sequences - sequences which were previously un-clustered or clustered into unusually small families. Further, using a simple merging strategy, we were able to merge 2,216 small families into 933 under-clustered families using the predicted missing sequences. Out of the 933 merged families, we could confirm correct mergings in at least 534 families using the tree-based over-clustering detection method. Finally, we provide the containerized versions of all the three methods that can be applied on any given set of gene families built using existing methods.

Introduction

Gene families, also known as orthologous groups, are groups of genes from a given set of species that have diverged from one another, from an ancestral gene in the most recent common ancestor of focal species. Gene families may contain genes that have diverged due to speciation and/or duplication. Accordingly, genes within a family may be classified as orthologs (separated by speciation) or in-paralogs (duplicated after the common ancestral node) [1–3]. For a number of analysis purposes -- for example, identification of candidates for drug/vaccine development [4–6] or annotation of newly sequenced genomes by cross-referencing function information from multiple species [7–11] -- it is useful to identify families such that all member genes originated from a single ancestral gene that was present in the common ancestor of the species under study. Many popular clustering techniques use these basic evolutionary properties of gene families for building gene families from whole proteomes of the species. These clustering algorithms use some form of normalized similarity/alignment scores between sequences as an input to a clustering method such as Markov Clustering (MCL) [12, 13] to generate gene family clusters.

Two of the most popular family building methods, OrthoFinder [14] and OrthoMCL [15], use normalized BLAST [16, 17] scores or E-values to cluster sequences into families using the chosen clustering algorithm. For Markov Clustering, which is widely used, the granularity of the MCL clusters is controlled by the inflation (I) parameter, with higher values of the parameter generally corresponding with a larger number of clusters. Both OrthoFinder and OrthoMCL use a single value of the Inflation parameter for building all families, for a given clustering run. Since different gene families can evolve at different rates [18–21], using a single Inflation parameter value for MCL clustering may be over-stringent for some families (resulting in fragmented/under-clustered families) - and under-stringent for other families (resulting in merged/over-clustered families).

In this study, we present a machine learning method for detection and correction of under-clustered families. The method evaluates sequence pairs, using a training set (deriving an appropriate family-specific homology threshold), and a test set. Potentially missing sequences for the family are predicted using family-specific alignment score cutoffs obtained in the training step. The training step consists of repetitive model building and testing where, during each iteration, a combined set of sequences from the given family along with a selected set of closest non-family sequences is randomly split into training and testing parts. The training part of the family is used to build a Hidden Markov Model (HMM) [22], which is then tested to recognize the correct family sequences from the testing part, containing both family and non-family sequences. Pairs of sequences were used as data points for training and testing the family models, and for inferring suitable per-family alignment score cutoffs.

We also present a method for comparing different family sets or clustering solutions, built using same set of species and a method for detecting potential over-clustering in families,

by scoring their rooted family phylogenies. Family consistency can also be assessed by comparing families generated by different family construction methods. This requires comparing two sets of families to establish family correspondences between the two sets. The family correspondences are established by examining the two-way family overlaps between the sets. Degrees of overlaps between corresponding families, between the two sets, can be used as measure of family consistency.

Potential over-clustering can be detected by comparing gene family phylogenies to known species phylogenies to identify the proportion of ingroup sequences diverging after outgroup sequences. Since gene families contain sequences that have diverged at or after the divergence of the earliest diverging species under study, genes within families have also diverged after the divergence of outgroup species. We used this evolutionary property of gene families to score family trees for potential over-clustering.

The machine learning method was tested on curated set of yeast families, obtained from the YGOB database [23]. We take these as gold-standard ("true") families, for comparison. We also generated an intentionally under-clustered set of yeast families to check the ability of the method to detect under-clustered families. The family-sets comparison method was used to compare the OrthoFinder-derived yeast families to the curated reference set of families, in order to detect under-clustered families produced by OrthoFinder. The machine learning method was applied on these under-clustered OrthoFinder families, to demonstrate the ability of the method in correcting under-clustered families. We also used the machine learning algorithm for merging small legume families into larger families (generated with OrthoFinder), to demonstrate the ability of the method in merging fragmented families. We also used the tree-based over-

clustering detection method to score and evaluate merged families, in order to check the correctness of the mergings produced by the methods above.

Methods

Under-clustering Detection and Correction

Collecting candidate missing sequences

Sequences from a given family were searched against the database of proteomes from all the species under study. With each family sequence as query, all the hits that match the query better than the worst-matching family hit were collected into a list. Each query sequence can attract one or more non-family sequences along with the original family sequences. A combined list of sequences was prepared from the lists of hits collected from searching all the family sequences against the database (Fig 2.1). This combined list contains all of the original family sequences and can contain one or more non-family sequences. The non-family sequences in the list of retrieved sequences are candidates for sequences missing from the family.

The phmmer program from the HMMER package (version 3.1b2) [24] was used for searching families against the database of whole proteomes. E-values from phmmer output were used to rank the hits, in order to find the worst-matching family hit.

Defining sequence pairs

The list of retrieved sequences obtained in the previous step was used to define family and non-family sequence pairs (Fig 2.2). Family pairs are those exclusively between original family members, and non-family pairs are those between family and non-family members. Family pairs were labeled as “positive” and non-family pairs were labeled as “negative”.

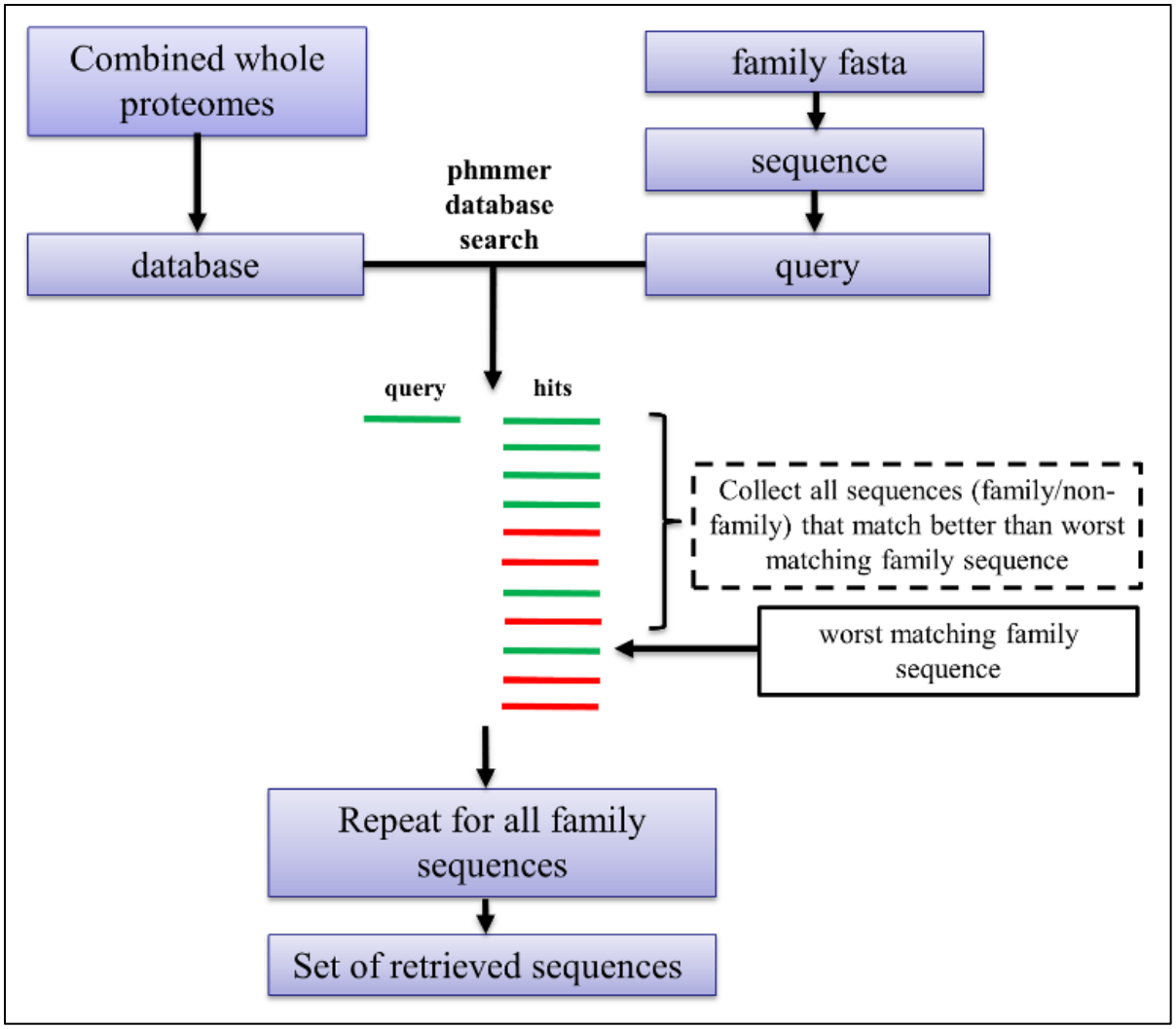

Fig 2.1. Procedure for collecting candidate missing (non-family) sequences for a given family. Each sequence from a family was searched against the proteomes from target and outgroup species using the phmmer program. For every query sequence, all the hits that match the query with better scores than the worst-matching family sequence were added to a set of retrieved sequences containing all the original family sequences, plus the closest non-family sequences. The non-family sequences are candidates for sequences missing from the original family.

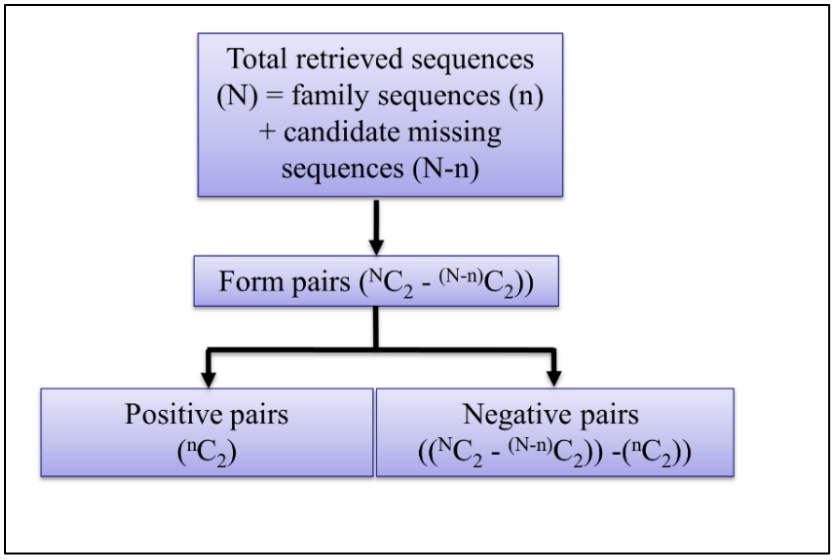

Fig 2.2. Defining positive and negative pairs from the family and non-family sequences obtained in the previous step. Let 'N' be the total number of sequences in the list of retrieved sequences for a given family and let 'n' be the number of original family sequences. Then, (N-n) is the number of non-family sequences in the retrieved list of sequences that could be missing from the family. Pairs are only formed between the family members (positive pairs) and between the family members and non-family sequences (negative pairs). The number of pairs that can be formed from a set of retrieved sequences containing 'N' sequences with 'n' sequences from the original family is $({}^{N}C_2 - {}^{(N-n)}C_2)$. From these pairs, there are ${}^{n}C_2$ positive pairs and $(({}^{N}C_2 - {}^{(N-n)}C_2) - {}^{n}C_2)$ negative pairs.

Training and classification statistics

HMM-based pair-classification models were built to classify the positive (family) pairs from negative (non-family) pairs, for a given family, and 10 iterations of repeated test/train split strategy were used to assess the classification performance (Fig 2.3). For each iteration, the set of positive pairs was randomly split into training (80%) and test (20%) sets. The HMM model was

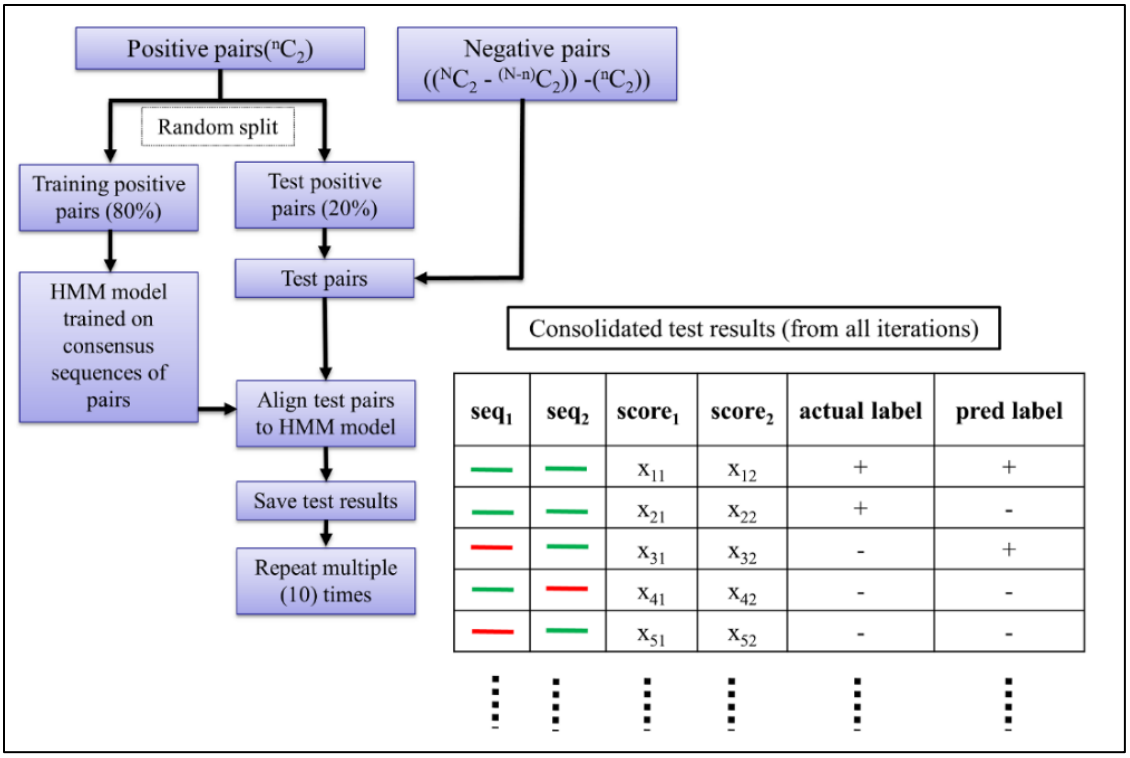

Fig 2.3. Model training and testing for classifying the positive pairs from the negative pairs.

For each iteration, the set of positive pairs was randomly split into training and testing sets, with the training set containing 80% of positive pairs and the testing set containing the remaining 20%. Consensus sequences of positive pairs from the training set were used to build HMM and the positive pairs from the test set and all the negative pairs, were aligned to the HMM. Individual sequences of each test pair were aligned to the HMM and full sequence alignment scores for both sequences were recorded. Alignment score results obtained from aligning unseen test pairs to trained HMMs, consolidated from all test-train split iterations performed on a given family, were used for calculating classification statistics for the family to analyze the separation between positive and negative pairs. The alignment scores of the individual sequences of the test pairs were used to predict the test pairs as positive or negative. Given a fixed score cutoff, a test pair was predicted as positive if both alignment scores corresponding to both the sequences of the pair were greater than or equal to a score cutoff, else the test pair was predicted as negative.

Accordingly, for a fixed score cutoff, True Positive (TP) test pairs were those that were originally positive and were also predicted as positive, False Positive (FP) test pairs were those that were originally negative but were predicted as positive, and finally, False Negative (FN) test pairs were those that were originally positive but were predicted as negative.

trained on consensus sequences of positive pairs from the training split, obtained using the "-c" option of hmmemit program (HMMER package:version 3.1b2). The MAFFT program (version v7.407) [25] was used for calculating the multiple sequence alignment used for building the HMMs. Subsequently, the trained HMM was tested on unseen positive pairs in the test split and the negative pairs by aligning individual sequences of the pairs to the HMM using the hmmsearch program. For each test pair, full sequence alignment scores for both the sequences were obtained. The test pair was predicted as positive if both the alignment scores were greater than or equal to a specified alignment score cutoff, else the test pair was predicted as negative.

The alignment scores for test pairs from all the iterations were consolidated (Fig 2.3) and used for calculating precision (TP/TP+FP), recall (TP/TP+FN) and F-score (eq 1) values for specified alignment score cutoffs, where TP, FP and FN are the number True Positive, False Positive and False Negative pairs, respectively. The F-score is defined as

$$F - score = (1 + \beta)\frac{precision \cdot recall}{(\beta^2 \cdot precision) \cdot recall} \qquad (1)$$

where β controls the importance of recall over precision with values greater than 1 favoring recall over precision [26]

Precision and recall values were obtained for a range of alignment score cutoffs, ranging from most to least stringent. Precision values were plotted against the corresponding recall

values to obtain the Precision-Recall curve (PR-curve) [27] and the area under the PR-curve (PR-AUC) was calculated using the trapezoidal rule [28]. The PR-AUC value ranges from 0 to 1 and was used as a score indicating family completeness. Complete families with no missing sequences are expected to have PR-AUC values closer to 1, indicating good separation between the positive and negative pairs. An example of PR-curve plot for a hypothetical gene family is shown in Fig 2.4. Each point on the curve corresponds to a recall value and the corresponding precision value (recall, precision) obtained using an alignment score cutoff with score cutoffs decreasing from left to right. The score cutoffs on the left produce pair classifications with high precision (low FP) but low recall (high FN). Conversely, low score cutoffs on the right produce pair classifications with low precision but high recall. An F-score can be calculated for each point (recall, precision) on the curve using the F-score function (eq 1). The point on the curve with the highest F-score is the point where optimal values of precision and recall exist (optimal trade-off between precision and recall). This point represents the best classification performance for the family and the corresponding score cutoff gives the best possible separation between the positive and negative pairs of the respective family.

Classification metrics such as the precision and recall values observed at the best F-score and the alignment score cutoff which gives the best F-score were calculated. Two types of alignment score cutoffs were reported corresponding to the two values of the β parameter ($\beta = 1$ and $\beta = 2$) in the F-score function. The F-score function with $\beta = 1$ is called the F_1-score function and the F-score function with $\beta = 2$ is called the F_2-score function. In addition, the lowest alignment score observed for the positive pairs was also reported as the lowest alignment score cutoff for the given family.

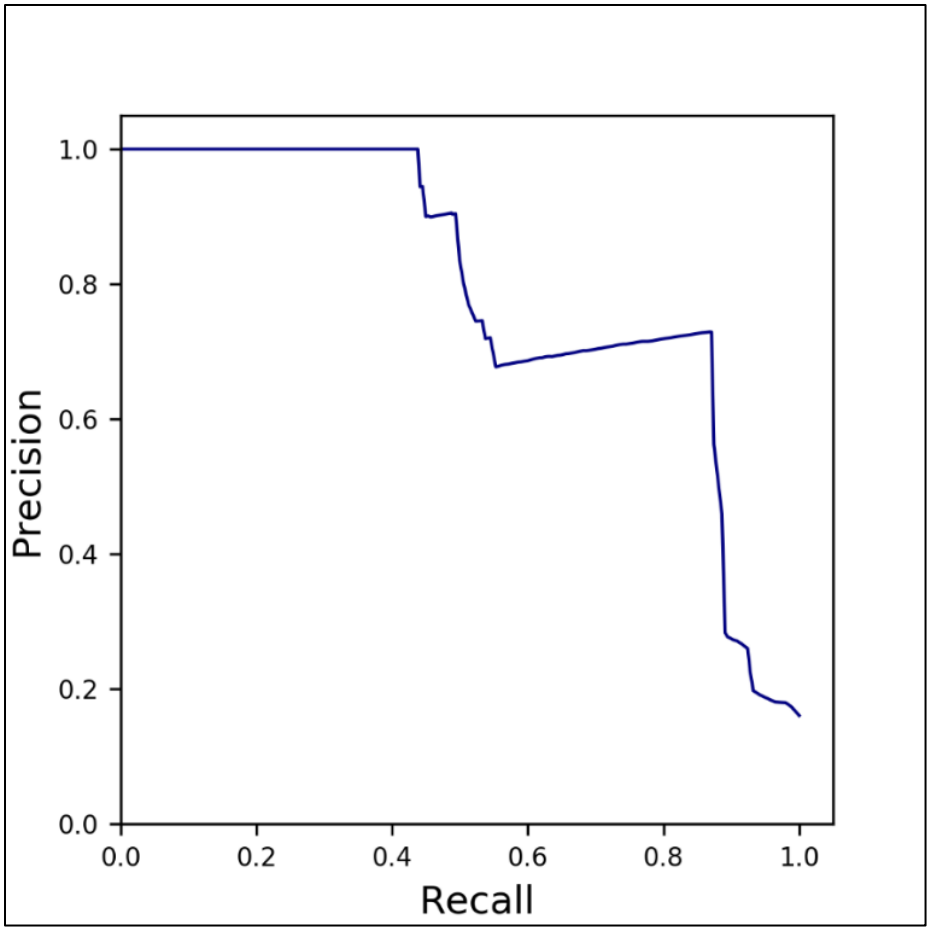

Fig 2.4. A typical Precision-Recall (PR) curve for a gene family. The precision and recall values obtained using different alignment score cutoffs were plotted against each other to obtain the Precision-Recall curve (PR-curve). The alignment score cutoffs vary from high value to low value from left to right, with cutoffs on the left giving high precision but low recall and cutoffs on the left giving low precision and high recall, for classification between the positive and negative pairs of a given family. The area under the PR-curve (PR-AUC) was calculated using the trapezoidal rule.

Predicting missing sequences

Missing sequences for every family were predicted using the negative pairs (Fig 2.5). A single HMM was built using the consensus sequences of all the positive pairs. Subsequently, all the negative pairs were aligned to this HMM and those negative pairs where the alignment scores for both the sequences of the pairs were greater than or equal to the chosen type of score cutoff (F1/F2/lowest) were re-classified as positive pairs. Unique sequences within these re-classified positive pairs were predicted and reported as the missing sequences for the family. The precision and recall values for prediction of missing sequences were also calculated, as (TP/(TP+FP)) and (TP/(TP+FN)), respectively, where True Positive (TP) are those predicted missing sequences that were truly missing from the family, False Positive (FP) are those that were predicted as missing but do not actually belong to the family and False Negative (FN) are those that are truly missing but were not predicted as missing.

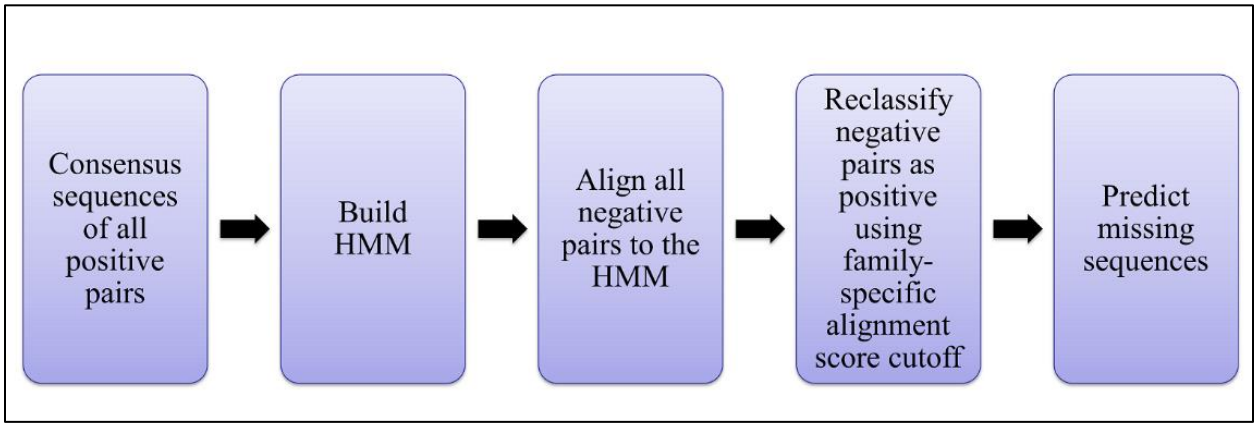

Fig 2.5. Predicting missing sequences using the negative pairs and family-specific alignment score cutoff. For each family, consensus sequences of the positive pairs and the negative pairs were aligned to the trained HMM for the family. Negative pairs where alignment scores of both the sequences of the pair were greater than or equal to the family-specific alignment score cutoff

were reclassified as positive pairs, and the constituent sequences were predicted as missing sequences.

Comparison of Family Sets

The family-sets comparison method was used to compare the reference set of yeast families to the yeast families rebuilt using OrthoFinder. For each family in both sets, the largest overlapping family in the opposite set is obtained. If, for example, fam1 from set1 is the largest overlapping family for fam2 from set2 and vice-versa, then fam1 and fam2 are considered corresponding families from both the sets. The two overlap scores, fam1 - fam2 overlap and fam2 - fam1 overlap are also calculated where fam1- fam2 overlap score is the proportion of sequences in fam1 that overlap with fam2 and, similarly, fam2 - fam1 overlap score is the proportion of sequences in fam2 that overlap with fam1. If both the scores are 1.0 for a pair of families, then the two families from the two sets match exactly.

Tree-based Over-clustering Detection

A rooted tree-based family scoring procedure was used to assess mergings and potential over-clustering in merged families obtained using the machine learning pair-classification method. Rooted phylogenies were built for each given family, together with the closest outgroup sequences, and analyzed for the presence of monophyletic ingroup clades. First, the closest outgroup sequences were identified for each family by searching the family HMM against the database of outgroup sequences and selecting the top 10 best matching outgroup sequences that align to the family HMM with e-value $\leq 10^{-5}$. Then, phylogenies were inferred for the combined set of family and outgroup sequences and rooted using the closest outgroup sequence. The RAxML [29] tool was used for construction Maximum Likelihood (ML) family phylogenies

with the PROTGAMMAAUTO substitution model, and were rooted using the closest available outgroup species

In the next step, quantitative scores were assigned to the rooted family phylogenies in order to reflect the number of monophyletic ingroup clades present in the trees. This scoring scheme is based on the proportion of ingroup sequence pairs that appear to diverge after the divergence of outgroup sequences. For a given rooted family tree, each pair of ingroup sequences found within the tree was labeled as True Positive (TP) or False Positive (FP) depending upon whether the pair appears to have diverged after or before the divergence of one or more outgroup sequences. The divergence status of any ingroup sequence pair in the tree was checked using the Most Recent Common Ancestor (MRCA) of the pair. All sequences corresponding to leaf nodes under this MRCA were collected and checked for the presence of one or more outgroup sequences. If no outgroup sequences were detected under the MRCA of an ingroup sequence pair, the pair is labelled as a TP, else it is labelled as FP. The FP label for any ingroup sequence pair indicates that there is at least one outgroup sequence that has diverged after the divergence of the pair and the corresponding sequences of the pair have been wrongly clustered into one family. A score for the family was calculated as TP/(TP+FP) which gives the proportion of ingroup sequence pairs diverging after the outgroup separation in the family tree.

Results

Behavior of the Machine Learning Method on “True” YGOB Families

The machine learning pair-classification-based scoring was tested on 4,796 yeast families from the Yeast Gene Order Browser (YGOB) database [23]. Since the YGOB families are built through manual curation using synteny-based evidence, we took them as correct or "true." To check if the machine learning method is assigning high classification performance scores to all

the true families, the distribution of the PR-AUC values for all the 4,796 yeast families was obtained (Fig 2.6A). As expected, the distribution is highly skewed towards PR-AUC value = 1.0, with 92% of family classifiers having values ≥0.75. This shows that the proposed method correctly recognizes complete families and assigns high classification performance scores to them.

Using the Machine Learning Method to Detect "Pure" but Under-clustered Families

Even though this under-clustering assessment method performs well in our tests on good/true families, it is important to study the behavior of the method on incomplete families. To check the behavior of the method on incomplete families, the "true" yeast families were modified so that each family is missing a random 20% of the family sequences.

The distribution of PR-AUC values for 4,796 artificially manipulated families where every family is missing 20% of their sequences is shown in Fig 2.6B. The distribution indicates that the machine learning performance of the families drops significantly, which shows the ability of proposed method to detect incomplete families. For 3,971 out of 4,796 families (83%), the PR-AUC value dropped significantly: PR-AUC ≥ 0.9 for true families and PR-AUC < 0.75 after removing family sequences (Fig 2.6C).

For each pure under-clustered family, the missing sequences were predicted back using lowest alignment score cutoff obtained during training. Since these incomplete families are "pure" i.e. they do not contain any non-family sequences, the lowest score cutoff can be regarded as a lower bound of the family. Any true family sequence/sequence pair is expected to align to the family HMM with a score greater than the lowest score cutoff. The predicted missing sequences were compared to the true missing sequences, for each family, and precision and recall values were calculated to study the accuracy of prediction. Out of 4,796 families, the

prediction performance for missing sequences was high (precision ≥ 0.75 and recall ≥ 0.75) for 3760 (78.4%) families with overall mean precision = 0.928 and overall mean recall = 0.859.

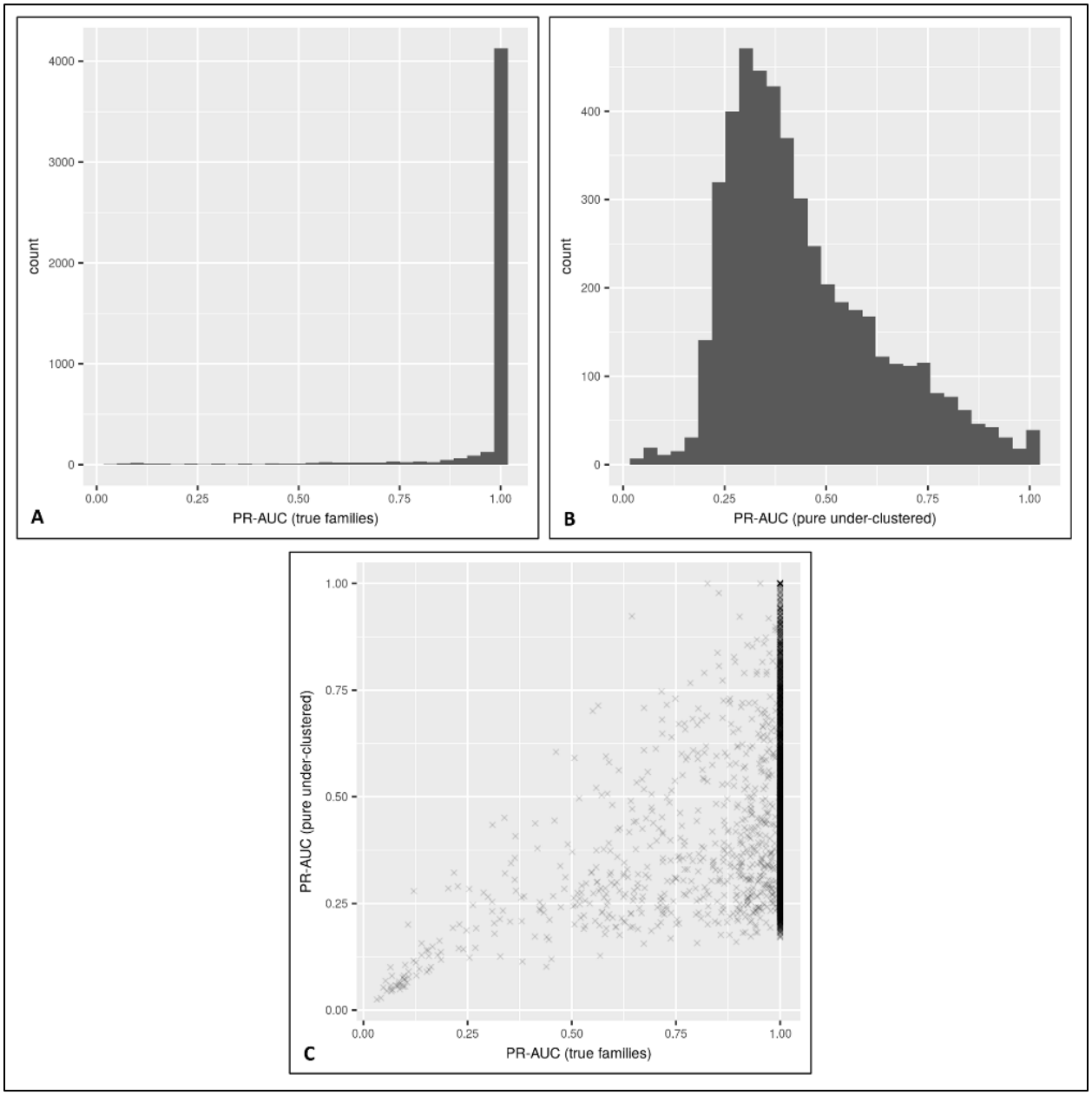

Fig 2.6. (A) Distribution of PR-AUC values for "true" yeast families. (B) Distribution of PR-AUC values for "pure under-clustered" yeast families with 20% of family sequences removed. (C) Scatterplot comparing the PR-AUC values of true families vs. incomplete families with 20% of sequences deleted. The distribution for true families is heavily skewed

towards 1.0 which shows that separation between the positive and negative pairs is good and in majority of the cases perfect, for true gene families. The perfect classification signifies that the families are complete with no missing sequences. The distribution for pure under-clustered families shows the drop in the PR-AUC values as compared to the distribution of PR-AUC values for true families. The scatterplot shows the drop in the PR-AUC value for each family after 20% of the sequences are removed from the family. Each point in the plot represents a family.

Application of the Machine Learning Method to Detect and Correct "Impure" and Under-clustered Families

Under-clustered families can also contain non-family or "wrong" sequences. To evaluate the behavior of our method on incomplete families contaminated with unrelated sequences, 2,391 yeast families were modified so that each family contained an additional 20% of sequences, from a set of the closest non-family sequences - in addition to missing 20% of the original family sequences. We analyzed these "impure, under-clustered families" using the machine learning method. Since these under-clustered families contain unrelated sequences, it is possible that more unrelated sequences could be attracted while selecting the candidate missing sequences, in the first step of the workflow (See methods). To make sure only relevant missing candidates are selected, only those non-family sequences were retained in the first step that were attracted by at least 50% of family sequences.

As observed in the case of pure under-clustered families, there was a significant reduction in the PR-AUC values for 83% of impure, under-clustered families, as compared to the true families, indicating that the machine learning method is able to detect under-clustering even when there are wrong sequences present in the under-clustered family. Since these families

already contain wrong/unrelated sequences, the lowest score cutoff that gives highest recall cannot be used for predicting missing sequences. Therefore, two different types of alignment score cutoffs - score cutoff obtained using the F_1-score function and score cutoff obtained using the F_2-score function, were used to predict the missing sequences for each of the 2,391 impure under-clustered families. Table 2.1 shows the precision and recall results for prediction of missing sequences obtained using the two types of score cutoffs.

Table 2.1. Precision and recall results for predicting missing sequences, for yeast families containing non-family sequences in addition to missing 20% of the sequences

Function type	**% of families with prediction precision ≥0.75**	**% of families with prediction recall ≥ 0.75**	**% of families with both prediction precision ≥ 0.75 and prediction recall ≥ 0.75**	**Mean prediction precision**	**Mean prediction recall**
F1-score	1729/2391 = 72.3	1348/2391 = 56.3	1138/2391 = 47.5	0.766	0.666
F2-score	1685/2391 = 70.4	1978/2391 = 82.7	1426/2391 = 63.1	0.790	0.868

The prediction results show that the alignment score cutoffs obtained using the F_1-score function predicts missing sequences with high precision for 72% of the families but fails to recognize true missing sequences for a majority of the families, with only 56% of the families having high recall. Consequently, the overall prediction performance is high (high precision with high recall) for only 47% of the families with mean precision = 0.766 and mean recall = 0.666. In order to increase the recall performance, score cutoffs obtained using F_2-score were also used to predict missing sequences for same set of impure under-clustered families. The F_2-score function is expected to give alignment score cutoffs that favor recall over precision. Accordingly,

the alignment score cutoffs obtained using the F_2-score function improved the recall for prediction of missing sequences with 82% families having high recall and with 70% families having high precision performance with mean precision = 0.790 and mean recall = 0.868. This has also increased the overall prediction performance with 63% of families having high prediction precision and recall for predicting missing sequences.

Comparing Families Obtained from Existing Methods to Reference Families

The OrthoFinder method has been shown to outperform many popular family building methods like OrthoMCL, TreeFam and OMA [15, 30, 31]. Therefore, we used OrthoFinder to rebuild the yeast families from 20 yeast proteomes, from the YGOB database. Subsequently, we applied the family-sets comparison method to compare the yeast families built using OrthoFinder to the reference yeast families from the YGOB database. A total of 7,513 pairs of corresponding families were detected between the two sets, out of which, 6,085 families were perfectly rebuilt by OrthoFinder i.e. the two-way family overlaps were exactly equal to 1.0 (See Methods). Out of the remaining 1,428 family correspondences, for 422 family pairs, the reference – orthofinder overlap was less than 1.0 and orthofinder – reference overlap was equal to 1.0, which meant these families were under-clustered by OrthoFinder. Similarly, for 875 family pairs, out of 1,428, the reference – orthofinder overlap was equal to 1.0 and orthofinder – reference overlap was less than 1.0, which meant these families were over-clustered by OrthoFinder. Finally, for 131 family pairs, both overlaps were less than 1.0, which meant that these families were both under-clustered and over-clustered by OrthoFinder.

Application of the Machine Learning Method to Improve Families from Existing Family Building Methods

To check if the machine learning method can improve families built using existing family building methods, we analyzed 374 under-clustered yeast families from OrthoFinder. The machine learning method was able to improve 5 to 19 under-clustered families by predicting missing sequences using the family-specific alignment score cutoffs obtained using F_1-score and F_2-score functions, respectively. Table 2.2 shows the results on improving the under-clustered OrthoFinder families using two types of alignment score cutoffs obtained using the two F-score functions.

Table 2.2. Results on correcting under-clustered yeast families obtained using OrthoFinder.

Function Type	No. of improved families	Mean prediction precision	Mean prediction recall
F_1-score	5	0.833	0.480
F_2-score	19	0.904	0.648

As expected, the more conservative F_1-score function corrects fewer families: only 5, with high precision of 0.833 and low recall of 0.48. In comparison, the F_2-score function improves 19 incomplete families, with significantly more recall of 0.65.

Analyzing and Correcting OrthoFinder Legume Families

We also analyzed 14,663 legume families built using the OrthoFinder (version 2.2.0) tool from 14 legume proteomes [32–43] (Table 2.3), using the machine learning method. The 14 legume species belong to subfamily Papilionoideae of family Fabaceae (the third largest family of flowering plants [44]). An ancient Whole Genome Duplication (WGD) occurred in the

common ancestor of the Papilionoid sub-family, around 55 Ma [45–52]. In addition, some genera such as *Glycine* and *Lupinus* have also undergone independent, lineage specific WGDs [52, 53].

Table 2.3. Genome and annotation sources and versions.

Species	Genotype	Assembly	Annot.	Publication	Source
Arachis duranensis	V14167	1	1	Bertioli et al. (2015)	PeanutBase
Arachis ipaensis	K30076	1	1	Bertioli et al. (2015)	PeanutBase
Arachis hypogaea				Bertioli et al. (2015)	PeanutBase
Cajanus cajan	ICPL87119	1	1	Varshney et al. (2012)	LegumeInfo
Cicer arietinum	Frontier	1	1	Varshney et al. (2013)	LegumeInfo
Glycine max	Williams 82	2	1	Schmutz et al. (2010)	Phytozome
Lotus japonicus	MG20	3	1	Sato et al. (2008)	Phytozome
Lupinus angustifolius				Hane et al. (2016)	LegumeInfo
Medicago truncatula	A17_HM341	4	2	Tang et al. (2014)	Phytozome
Phaseolus vulgaris	G19833	2	1	Schmutz et al. (2014)	Phytozome
Trifolium pratense				De Vega (2015)	LegumeInfo
Vigna angularis	Va3.0	1	3	Kang et al. (2015)	LegumeInfo
Vigna radiata	VC1973A	6	1	Kang et al. (2014)	LegumeInfo
Vigna unguiculata	IT97K	1	1	Phytozome	Phytozome

The distribution of family sizes up to size 100 is shown in Fig 2.7. Approximately 12% of sequences (64,047) were not clustered by OrthoFinder (i.e. remained as singletons), and there were 10,963 families with 2 to 8 sequences (small and potentially under-clustered).

Our hypothesis is that these small families and the unclustered sequences could be a result of over-fragmentation or under-clustering of larger families, due to stringent clustering parameters. Accordingly, 14,663 larger families (with sizes between 9 and 36) were analyzed using the machine learning method. The unclustered sequences and sequences from smaller

families (with 2 to 8 sequences) were compiled together to be used as the sequence database for phmmer searches, for gathering candidate missing sequences from the 14,663 families. Only

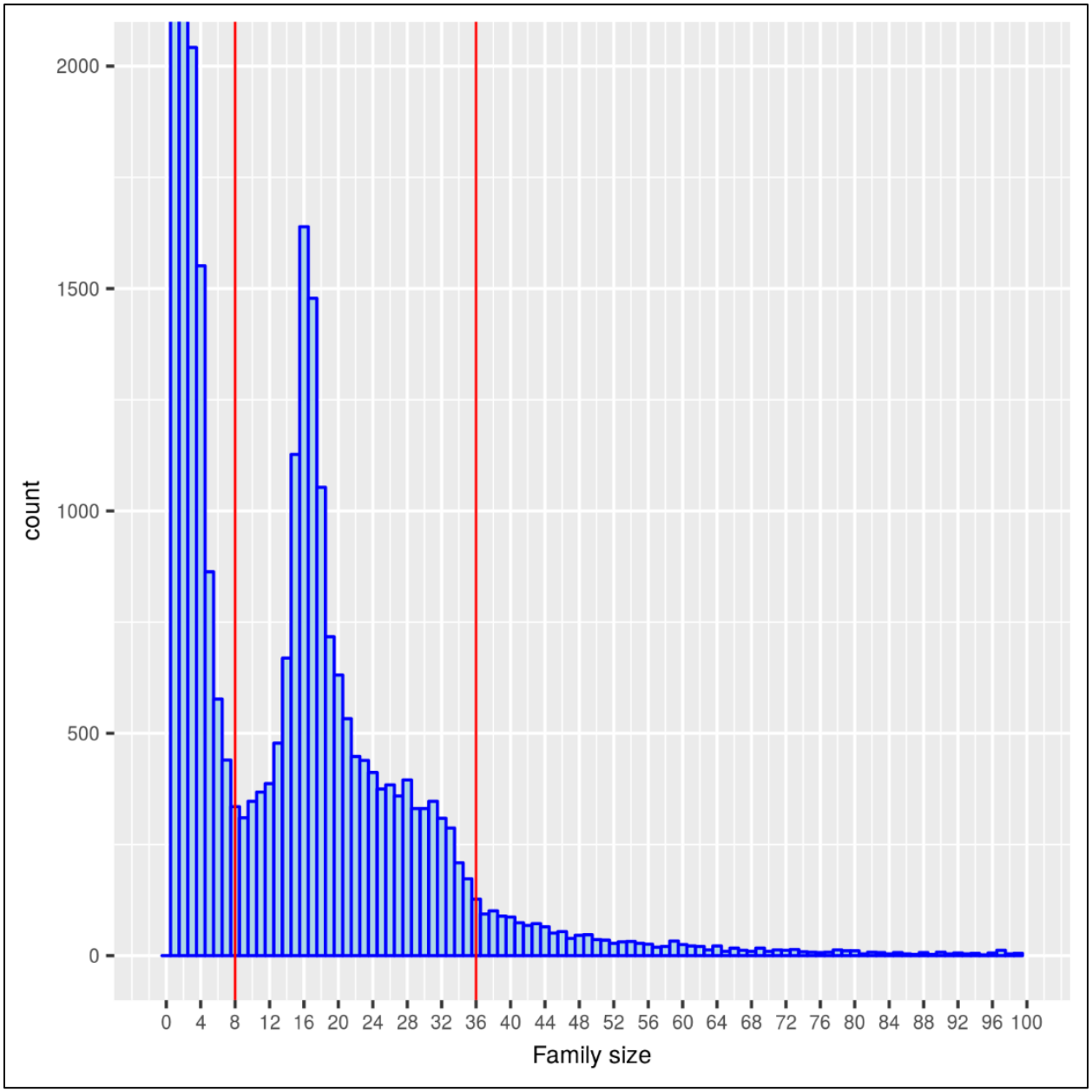

Fig 2.7. Family size distribution for legume families built using OrthoFinder. The size distribution is shown up to size 100. The families with sizes between 1 and 8 were considered unusually small. The families that fall between the vertical red lines ($9 \leq \text{size} \leq 36$) were analyzed using the machine learning method.

those non-family sequences were selected as candidate missing sequences that were attracted by at least 50% of family sequences. For 9,581 of the 14,663 families, no candidate missing sequences were found according to this selection criteria.

For the remaining 5,082 families for which one or more candidate missing sequences were detected, machine learning models were built and tested to study the separation between the positive pairs formed within the family sequences and the negative pairs formed between the family and non-family sequences. The distribution of PR-AUC values for the 5,082 legume families is shown in Fig 2.8. The distribution is skewed towards 1.0, showing that most of these families have good separation between the positive and negative pairs.

For 1,665 families out of 5,082, one or more missing sequences were predicted using the family-specific alignment score cutoffs obtained in training, using the F_1-score function. The F_1-score function was used to obtain the optimal alignment score cutoffs, as family precision was considered more important than recall for predicting the missing sequences. There were 3,588 sequences from the small families that were predicted as missing from 1,665 larger families.

Next, we attempted to merge the smaller families into the larger families using the predicted missing sequences with the following merging rule. If, for a larger family, missing sequences were predicted from a smaller family that were more than 50% of the size of the smaller family, the corresponding smaller family was merged into the larger family. For small families that had potential to merge into more than one larger family, the hhsearch program (version 2.0.16) [54], from the hh-suite package, was used to select the best-matching larger family. In all, 2,216 small families were merged into 933 larger families using the machine learning method.

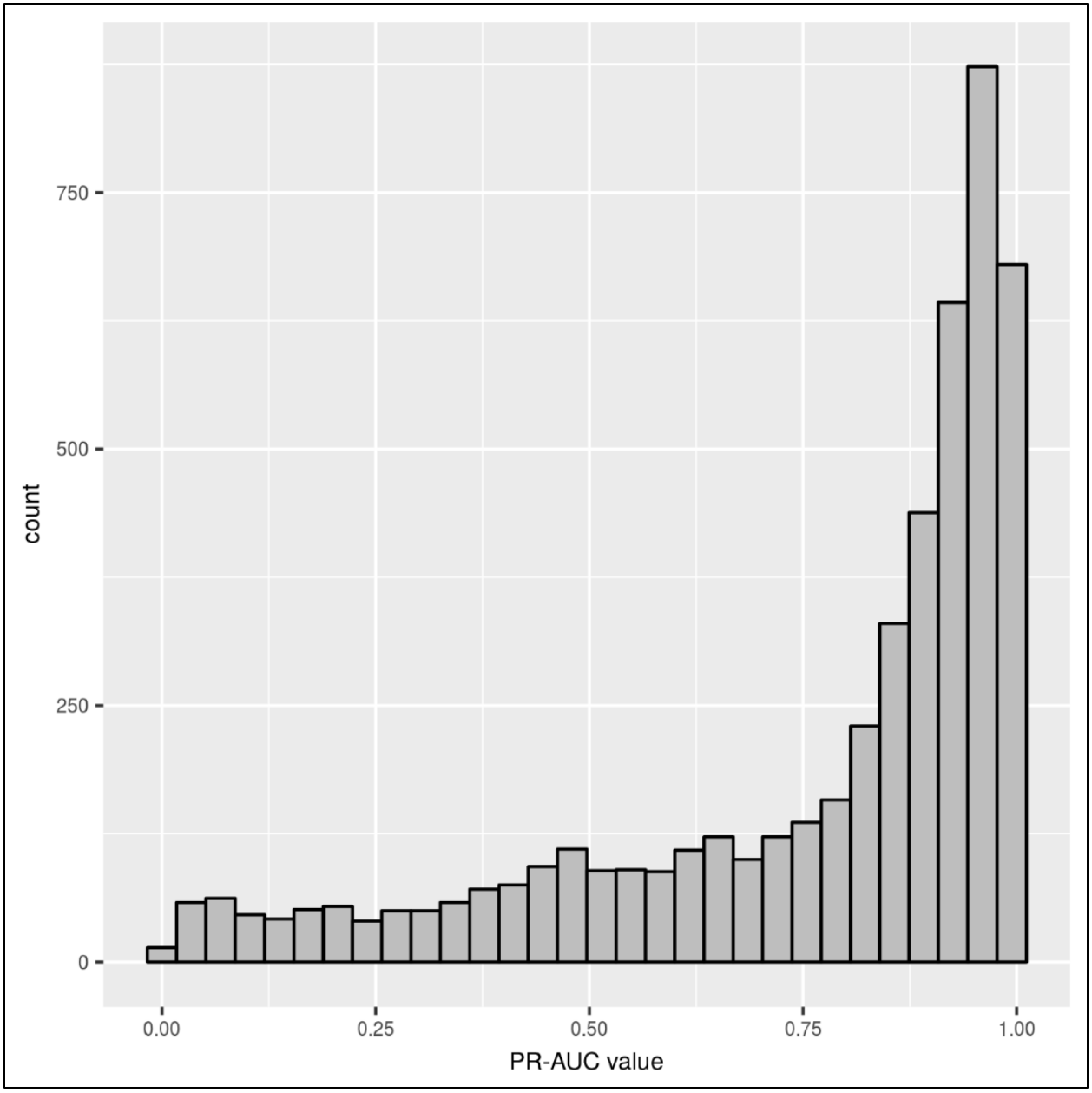

Fig 2.8. Distribution of PR-AUC values for 5082 families analyzed using the under-clustering detection and correction method. The distribution is skewed towards higher PR-AUC values, signifying good classification performance for the majority of the families.

We used the tree-based over-clustering detection method to check the correctness of the family mergings. The tree-based scoring method was applied to each of the 933 merged families for detecting presence of monophyletic clades containing all the legume sequences, with respect to legume outgroups. The merging was considered correct if only one monophyletic legume clade was observed in the merged family tree, i.e. the tree score was equal to 1. Out of the 933 trees, the tree score was equal to 1 for 478 families. Also, there were 56 families with tree scores less than 1 and containing more than one legume clades, but the newly merged sequences were part of a major clade containing 70% or more sequences from the merged family, and the minor clades contained 30% or less sequences that were part of the original unmerged family. For at least 534 families out of the 933 corrected families, family mergings predicted by the machine learning method were consistent with expected phylogenetic relationships.

We also attempted to predict missing sequences for the 14,633 legume families using a simple HMM searching strategy, to highlight the importance of family-specific alignment score cutoffs. Every family HMM was searched against the database of sequences from the small families and sequences that align to the family with e-value $\leq 1e^{-10}$ were predicted as missing sequences for the family. This resulted in unusually large family clusters, with more than 1600 clusters containing ≥ 100 sequences, and the largest clusters containing more than 1,400 sequences. This shows that the generic E-value cutoff is too relaxed for some families. For example, for family OG0001825 (size = 36), 6 sequences were predicted as missing using the family-specific alignment score cutoff through the machine learning method, as opposed to 227 sequences that were predicted as missing using the simple HMM searching strategy with a e-value cutoff of $1e^{-10}$.

Discussion

Under-clustering and over-clustering are common problems in current family building methods. For example, at least 374 under-clustered and 875 over-clustered yeast families were produced by the OrthoFinder method. In the case of legume families, the OrthoFinder method could not assign about 12% of the genes to any family, and many small families were also produced - potentially indicating fragmentation of larger families. In this work, we present methods for detection and correction of under-clustered and over-clustered families using a sequence-pair-based classification approach, a family-sets comparison approach, and a tree-based family scoring approach.

The machine learning method was tested on "true" and modified yeast families to check the effectiveness of the methods in detecting complete families and in detecting and correcting under-clustered families. On the true yeast families, the method correctly identified complete families, assigning near-optimal or perfect PR-AUC values to the unmodified families, and also identifying under-clustered families through low or sub-optimal values for the PR-AUC statistic. These results show that the family-specific alignment score cutoffs obtained during training the machine learning models were able to recognize true missing sequences for the families even when unrelated sequences were present in the families.

The family-sets comparison method was used to compare the reference set of yeast families, obtained from the YGOB database, to the set of families built using OrthoFinder. Out of the total number of yeast families produced by OrthoFinder, the family-sets comparison method was able to detect at least 374 yeast families that were under-clustered. To check if the machine learning method can improve these families, we applied the machine learning method, finding that it was able to improve up to 19 families by predicting the correct missing sequences for these families, with mean precision of 0.9 and mean recall of 0.65.

Finally, we also applied the machine learning method for analyzing 14,633 legume families built using OrthoFinder. The machine learning method was able to identify 3,588 missing sequences for 1,665 families, which were subsequently used to merge 2,216 small families into 933 larger families using a simple merging rule. The tree-based over-clustering detection method used to score and analyze the phylogenies of the merged families provided confirmatory evidence for correct mergings in at least 534 of the 933 merged families.

The machine learning based under-clustering detection and correction method can employ different types of family-specific alignment score cutoffs for predicting missing sequences, depending upon the nature of under-clustering and preference of family precision or family completeness. There is a tradeoff between the objectives of family accuracy and family completeness. If family precision is valued more highly than family completeness, then an alignment score cutoff with high value is recommended. Conversely, if family completeness (recall) is preferred over precision, then a low alignment score cutoff should be used for predicting missing sequences. The alignment score cutoff obtained using the F_1-score function (i.e. F-score with $\beta = 1$) appears to be predicting missing family sequences with high precision and low recall, as seen from the prediction results on impure under-clustered families. Therefore, the alignment score cutoff obtained using the F_1-score function can be used to predict missing sequences for families with high precision. On the other hand, the alignment score cutoffs obtained using the F_2-score function improves the recall for prediction at the expense of precision, favoring recall over precision. The high-recall alignment score cutoff obtained using the F_2-score function can be used for predicting missing family sequences with more recall at the expense of precision. The lowest alignment score cutoff with the highest possible recall can be used in case of those under-clustered families that are not “contaminated” with unrelated

sequences. Based on the results obtained from modified under-clustered yeast families, we can make the following recommendations on which type of alignment score cutoff to use for a given set of families. If the user/expert is confident that any family detected as under-clustered (has a low PR-AUC value) has a low probability of containing non-family/unrelated sequences, then the score cutoff obtained using the F_2-score or the lowest possible score cutoff can be reliably used for predicting the missing family sequences. In contrast, if the user thinks that the under-clustered family may contain up to 20% of unrelated sequences, then the more conservative score cutoff obtained using the F_1-score function should be used for predicting the missing sequences.

Although the machine learning method shows good results empirically, we note a statistical weakness in the method that should be considered when interpreting results. Because the method trains the HMMs and tests on pairs of sequences obtained from a family, there may be overlap of information between the training and testing sets, due to overlap of sequences between the sets. For example, if a family contains three sequences: A, B and C, and the training is performed using the sequence pairs A-B and B-C with the test set containing the sequence pair A-C, there is overlap of information between the training and the test sets since sequences A and C are present in both the sets, even though training was performed on consensus sequences of individual pairs and not the sequences themselves. This overlap of information between the training and test sets might create family models that are biased towards recognizing and family sequences and biased against accepting true missing sequences. This bias might result in overestimation of family-specific alignment score cutoffs i.e. score cutoffs that are higher or more stringent than the “true” score cutoffs for the given families.

The methods presented here can be used as a post-processing tools for independently assessing gene family sets built using existing family building methods. Every family can be analyzed using the methods for signs of under-clustering or over-clustering, using the machine learning, family-sets comparison and tree-based over-clustering detection methods. We also provide the containerized versions of all the tools presented in this work. The container for the machine learning method for detection and correction of under-clustered families can be downloaded from https://hub.docker.com/r/akshayayadav/undercl-detection-correction. The containers for the family-sets comparison method and the tree-based over-clustering detection method can be downloaded from https://hub.docker.com/r/akshayayadav/family-sets-comparison-tool and https://hub.docker.com/repository/docker/akshayayadav/overcl-detection-correction, respectively.

References

1. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

2. Fitch WM (2000) Homology: a personal view on some of the problems. Trends Genet 16:227–231

3. Sonnhammer EL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. TRENDS Genet 18:619–620

4. Kissinger JC, Brunk BP, Crabtree J, Fraunholz MJ, Gajria B, Milgram AJ, Pearson DS, Schug J, Bahl A, Diskin SJ (2002) The Plasmodium genome database. Nature 419:490

5. Whetzel PL, Date SV, Gajria K, Fraunholz MJ, Gajria B, Grant GR, Iodice J, Labo PT, Milgram AJ, Stoeckert CJ (2005) PlasmoDB: The Plasmodium Genome Resource. In: Mol. Approaches Malar. American Society of Microbiology, pp 12–23

6. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419:498

7. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

8. Galperin MY, Koonin EV (1999) Searching for drug targets in microbial genomes. Curr Opin Biotechnol 10:571–578

9. Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV (2000) Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol 1:research0009.1

10. Natale DA, Galperin MY, Tatusov RL, Koonin EV (2000) Using the COG database to improve gene recognition in complete genomes. Genetica 108:9–17

11. Forterre P (2002) A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet 18:236–237

12. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584

13. Van Dongen SM (2000) Graph clustering by flow simulation. PhD Thesis

14. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

15. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res 13:2178–2189

16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

17. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

18. Ohta T (2000) Evolution of gene families. Gene 259:45–52

19. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. PloS One 1:e85

20. De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22:1269–1271

21. Trachana K, Larsson TA, Powell S, Chen W-H, Doerks T, Muller J, Bork P (2011) Orthology prediction methods: a quality assessment using curated protein families. BioEssays News Rev Mol Cell Dev Biol 33:769–780

22. Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6:361–365

23. Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15:1456–1461

24. Eddy S (2003) HMMER User's Guide. Biological Sequence Analysis Using Profile Hidden Markov Models.

25. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780

26. Rijsbergen CJV (1979) Information Retrieval, 2nd ed. Butterworth-Heinemann, Newton, MA, USA

27. Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: Proc. 23rd Int. Conf. Mach. Learn. ACM, pp 233–240

28. Atkinson KE (2008) An introduction to numerical analysis. John Wiley & Sons

29. Stamatakis A, Hoover P, Rougemont J (2008) A Rapid Bootstrap Algorithm for the RAxML Web Servers. Syst Biol 57:758–771

30. Li H, Coghlan A, Ruan J, Coin LJ, Heriche J-K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34:D572–D580

31. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2010) OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 39:D289–D294

32. Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, Liu X, Gao D, Clevenger J, Dash S (2015) The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat Genet 47:438

33. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83

34. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240

35. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J (2010) Genome sequence of the palaeopolyploid soybean. nature 463:178

36. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227–239

37. Hane JK, Ming Y, Kamphuis LG, et al (2017) A comprehensive draft genome sequence for lupin (Lupinus angustifolius), an emerging health food: insights into plant–microbe interactions and legume evolution. Plant Biotechnol J 15:318–330

38. Tang H, Krishnakumar V, Bidwell S, et al (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics 15:312

39. Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

40. De Vega JJ, Ayling S, Hegarty M, et al (2015) Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement. Sci Rep 5:17394

41. Kang YJ, Kim SK, Kim MY, et al (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. Nat Commun 5:5443

42. Kang YJ, Satyawan D, Shim S, et al (2015) Draft genome sequence of adzuki bean, *Vigna angularis*. Sci Rep 5:8069

43. *Vigna unguiculata* v1.1 (Cowpea). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Vunguiculata_er. Accessed 12 Feb 2019

44. Lewis GP (2005) Legumes of the World. Royal Botanic Gardens Kew

45. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16:1679–1691

46. Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome 47:868–876

47. Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ (2005) Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst Biol 54:441–454

48. Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. Proc Natl Acad Sci 103:14959–14964

49. Bertioli DJ, Moretzsohn MC, Madsen LH, Sandal N, Leal-Bertioli SC, Guimarães PM, Hougaard BK, Fredslund J, Schauser L, Nielsen AM (2009) An analysis of synteny of Arachis with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. BMC Genomics 10:45

50. Stai JS, Yadav A, Sinou C, Bruneau A, Doyle JJ, Fernández-Baca D, Cannon SB (2019) Cercis: A Non-polyploid Genomic Relic Within the Generally Polyploid Legume Family. Front Plant Sci.

51. Ren L, Huang W, Cannon SB (2019) Reconstruction of ancestral genome reveals chromosome evolution history for selected legume species. New Phytologist 223:2090–2103

52. Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart Jr CN, Rolf M (2014) Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol Biol Evol 32:193–210

53. Kroc M, Koczyk G, Święcicki W, Kilian A, Nelson MN (2014) New evidence of ancestral polyploidy in the Genistoid legume Lupinus angustifolius L.(narrow-leafed lupin). Theor Appl Genet 127:1237–1249

54. Söding J (2004) Protein homology detection by HMM–HMM comparison. Bioinformatics 21:951–960

CHAPTER 3. IMPROVING AND ANALYZING CURRENT LEGUME GENE FAMILIES

Akshay Yadav, David Fernández-Baca, Steven B. Cannon

Modified from a manuscript to be submitted to a peer reviewed journal

Abstract

The legume gene families at legumeinfo.org are built from 14 Papilionoid legume species, using methods that utilize differences in the synonymous-sites (K_s) in the gene sequences in order to capture the family clusters defined by the whole-genome duplication (WGD) that occurred in the Papilionoid ancestor. Here, we test methods for improving these gene families by detecting and merging fragmented or under-clustered families and splitting combined or over-clustered families, using HMM-based and tree-based methods, respectively. The family merging strategy is based on a two-way HMM-based database search procedure in which missing sequences are predicted for each family using their family HMMs and the outgroup sequences. Subsequently, a simple overlap rule is used to merge families using the predicted missing sequences. Using the two-way HMM-based search procedure, we were able to merge 1,720 families into 841 clusters. The same merging protocol was also used to reclassify 3,045 previously unclustered sequences into 347 families. A tree-based family splitting strategy was also applied to separate 2,554 merged and unmerged families in to 5495 families, that were detected as over-clustered by a sequence-pair-based family scoring method. We analyzed the improvements in the legume families after the application of merging and splitting procedures by comparing the protein domain compositions of the new families against the original families. We

provide the containerized versions of family merging, splitting and scoring methods along with the new set of improved legume families.

Introduction

The legume family (Leguminosae, Fabaceae) is the third largest family of flowering plants, comprised of approximately 750 genera and 20,000 species [1, 2]. The family originated around 59-64 million years ago (Mya) and rapidly diverged into 6 subfamilies [3, 4] with 4 of them - Papilionoideae, Caesalpinioideae, Detarioideae, Cercidoideae containing the majority of genera and species [2]. All four subfamilies have been shown to be affected by whole-genome duplications (WGDs), with especially strong evidence at the base of the Papilionoideae. For the other 3 lineages, the precise timing of WGDs remains uncertain due to low sampling [5]. Gene families hosted at legumeinfo.org in 2019 [6] are built from 14 legume proteomes all belonging to the subfamily Papilionoideae [7-18]. Table 3.1 lists the legume species and sources.

The families were built using a custom family construction method in order to circumscribe the family clusters such that each family captures all legume orthologs and paralogs deriving from speciations and legume WGDs, but not sequences deriving from earlier WGD events. The method used a combination of homology-filtering based on per-species synonymous site changes (K_s), comparisons with outgroup species, Markov clustering, and progressive refinements of family Hidden Markov Models (HMMs). The gene families are available at https://legumeinfo.org/data/public/Gene_families/legume.genefam.fam1.M65K/ and associated methods and scripts are available at https://github.com/LegumeFederation/legfed_gene_families.

Table 3.1. Genome and annotation sources and versions

Species	Genotype	Assembly	Annot.	Publication	Source
Arachis duranensis	V14167	1	1	Bertioli et al. 2015	PeanutBase
Arachis ipaensis	K30076	1	1	Bertioli et al. 2015	PeanutBase
Arachis hypogaea				Bertioli et al. 2015	PeanutBase
Cajanus cajan	ICPL87119	1	1	Varshney et al. 2012	LegumeInfo
Cicer arietinum	Frontier	1	1	Varshney et al. 2013	LegumeInfo
Glycine max	Williams 82	2	1	Schmutz et al. 2010	Phytozome
Lotus japonicus	MG20	3	1	Sato et al. 2008	Phytozome
Lupinus angustifolius				Hane et al. 2017	LegumeInfo
Medicago truncatula	A17_HM341	4	2	Tang et al. 2014	Phytozome
Phaseolus vulgaris	G19833	2	1	Schmutz et al. 2014	Phytozome
Trifolium pratense				De Vega et al. 2015	LegumeInfo
Vigna angularis	Va3.0	1	3	Kang et al. 2015	LegumeInfo
Vigna radiata	VC1973A	6	1	Kang et al. 2014	LegumeInfo
Vigna unguiculata	IT97K	1	1	Phytozome	Phytozome

Approximately 18k families were obtained using the custom K_s-based family building method, with family sizes ranging from 2 to 1260 sequences. Also, there were approximately 100k leftover/unclustered sequences that were not clustered in any of the families. The distribution of family sizes up to size 100 can be seen from Fig 3.1. The majority of the families fall between sizes 9 to 36 with two distinct peaks around sizes 16-20 and 28-32 corresponding to families that have lost the papilionoid WGD duplicate and to families that have retained the papilionoid WGD duplicate, respectively.

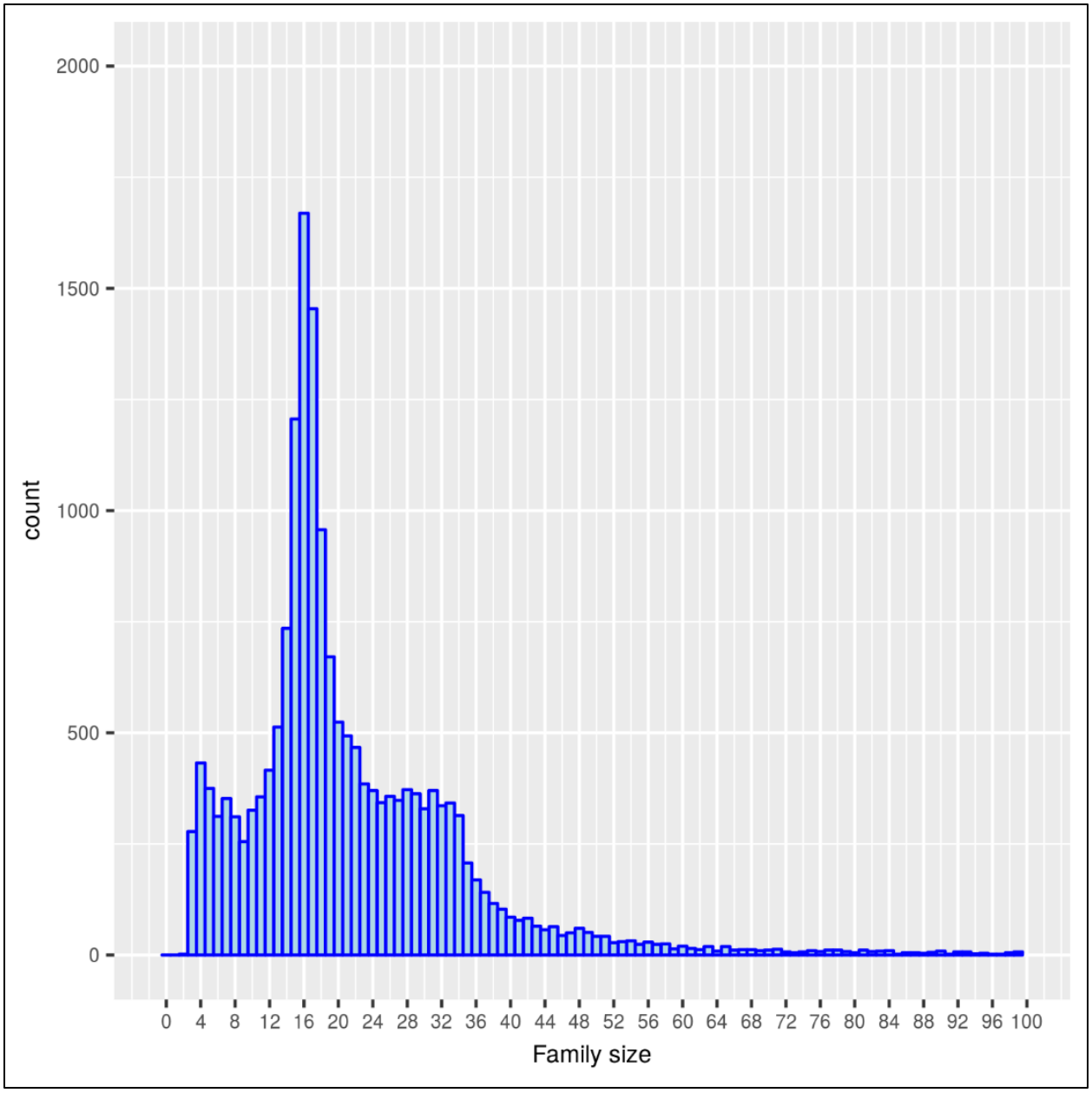

Fig 3.1. Size distribution of current legume families up to size 100.

In this study, we describe several general methods for assessing and improving the K_S-based legume families [6] using methods for detecting and correcting under-clustered and over-clustered gene families. Under-clustered gene families are those that are missing true sequences and could be produced due to fragmentation of larger families. Over-clustered families are those

that contain incorrect sequences due merging of two or more separate families. We used a Hidden Markov Model (HMM)-based searching method together with comparison to sequences from outgroup species for recognizing and merging under-clustered families and subsequently applied a tree-based method for detecting and splitting over-clustered families. We also employed a family scoring method based on protein domain composition within the families to study the change in the quality of the families with respect to the protein domains. Protein domains are sections of protein sequences that can fold and function independently. Consequently, multidomain proteins can evolve in a modular fashion through domain deletions, insertions or duplications in addition to sequence-based evolution [19]. Domain composition and/or content have been previously used to detect sequence homology with high accuracy [20, 21]. Closely related sequences from the same family can be expected to have fairly similar domain types and domain content. Accordingly, the domain-composition-based family scoring method assigns scores to families based on number of domains shared between the sequences of the family. This scoring scheme is expected to assign higher scores to families where most of the sequences have similar domain compositions. In general, well-conserved families are expected to have higher domain composition scores. Conversely, over-clustered families containing diverse sequences are expected to have lower scores due to less conservation of domain compositions across all the sequences.

Methods

HMM-based Family Merging

The *hmmsearch* and *hmmscan* programs from the HMMER package [22] are respectively designed for searching an HMM profile against a sequence database and searching individual

sequences against a database of HMM profiles. We first used *hmmsearch* to search each family HMM against the database containing sequences from all other legume families and sequences from outgroup species. All the legume sequences that align to the family better than any outgroup sequence and with e-value $\leq 10^{-5}$ were collected into a list called "closer sequences". We then took each sequence from the list of "closer sequences" to search against the database of all family HMMs and queries that find the original family as the best match were predicted as missing sequences for the family. The predicted missing sequences for each family were then used to merge families using the following merging rule: if for a family 'x', missing sequences were predicted from another family 'y' that were more than 50% of the size of 'y', then family 'y' was merged into family 'x'.

A similar strategy was used to reclassify previously unclustered sequences into families. In the *hmmsearch* step, each family HMM was searched against the database of unclustered sequences and outgroup sequences to find the candidate missing sequences; and in the next *hmmscan* step, the candidate missing sequences that find the corresponding family HMM as the best match among all the family HMMs were predicted as missing sequences for the family. Subsequently, new families were formed by including the previously unclustered sequences that were predicted as missing for the family. The mafft program [23] was used to generate Multiple Sequence Alignments (MSAs) for the families in order to build family HMMs.

Tree-based Family Scoring and Splitting

All the merged and unmerged families were subjected to a rooted-tree-based family splitting procedure where rooted phylogenies were built for each family, together with the closest outgroup sequences, and analyzed for the presence of monophyletic legume clades. First, the closest outgroup sequences were identified for each family by searching the family HMM

against the database of outgroup sequences and selecting the top 10 best matching outgroup sequences that align to the family HMM with e-value $\leq 10^{-5}$. Then, phylogenies were inferred for the combined set of family and outgroup sequences and rooted using the closest outgroup sequence. The FastTree tool [24] was used for construction Maximum Likelihood (ML) family phylogenies and functions from the ETE Toolkit [25, 26] were used to root the family trees.

In the next step, quantitative scores were assigned to the rooted family phylogenies in order to reflect the number of monophyletic legumes clades present in the trees. This scoring scheme is based on the proportion of legume sequence pairs that appear to diverge after the divergence of outgroup sequences. For a given rooted family tree, each pair of legume sequences found within the tree was labeled as True Positive (TP) or False Positive (FP) depending upon whether the pair appears to have diverged after or before the divergence of one or more outgroup sequences. The divergence status of any legume sequence pair in the tree was checked using the Most Recent Common Ancestor (MRCA) of the pair. All sequences corresponding to leaf nodes under this MRCA were collected and checked for the presence of one or more outgroup sequences. If no outgroup sequences were detected under the MRCA of a legume sequence pair, the pair is labelled as a TP, else it is labelled as FP. The FP label for any legume sequence pair indicates that there is at least one outgroup sequence that has diverged after the divergence of the pair and the corresponding sequences of the pair have been wrongly clustered into one family. A score for the family was calculated as TP/(TP+FP) which gives the proportion of legume sequence pairs diverging after the outgroup separation in the family tree.

Finally, the family trees that have suboptimal (<1.0) tree scores were split by separating individual monophyletic legume clades that contain at least 70% of the total legume species.

Python programs that utilize functions from the ETE Toolkit [25, 26] were used to process, score and separate the monophyletic legume clades from family trees.

Protein-domain-composition-based Family Scoring

Gene families were also scored to reflect the protein domain composition of their constituent sequences. Pfam [27] domains were assigned to all sequences within each family using the *pfam_scan.pl* [28] program. For each pair of sequences within a given family, domain feature vectors were defined for the sequences of the pair based on the combined set of domains detected in both the sequences. The cosine similarity score was calculated between the two feature vectors. For example, suppose sequences X and Y from the same family have the domain orderings, (A, B, B, C) and (A, A, B, D), respectively. The duplicate domains in both the sequences are assigned unique ids to distinguish them from each other, X: (A, B, B-2, C) and Y:(A, A-2, B, D). The domain content universe for both the sequences is (A, A-2, B, B-2, C, D). Accordingly, the domain feature vector for both sequences is X: (x_1, 0, x_2, x_3, x_4, 0) and Y: (y_1, y_2, y_3, 0, 0, y_4), where x_i and y_i are the alignment scores for the corresponding domain HMMs aligning against the sequences X and Y, respectively. Cosine similarities (eq 1) were calculated between all pairs of sequences using their domain feature vectors. The domain composition score for the family was calculated as the mean of all pairwise cosine scores.

$$C(X,Y) = \frac{\sum_i x_i y_i}{\sum_i x_i^2 \sum_i y_i^2} \qquad (1)$$

Results

The HMM-based protocol was applied on all ~18k K_s-based legume families. Sequences from 5 outgroup species: *Prunus persica*, *Cucumis sativus*, *Arabidopsis thaliana*, *Vitis vinifera*

and *Solanum lycopersicum* were used to collect the closer non-family sequences for each family (See Methods). Using the closer sequences, 32,402 missing sequences were predicted for 3,679 families. Subsequently, the predicted missing sequences were used for merging 1,720 families into 841 clusters using the 50% merging rule (See Methods). The same HMM-based searching protocol was also used to reclassify 3,045 previously unclustered sequences into 347 families. Fig. 3.2 shows the size distributions, up to family size 60, of the 1720 families that were merged to produce 841 clusters and the 347 families into which previously unclustered sequences were classified. Both the distributions show a majority involvement of small families in merging and accepting unclustered sequences. We see 2 distinct peaks in the size distribution of merged

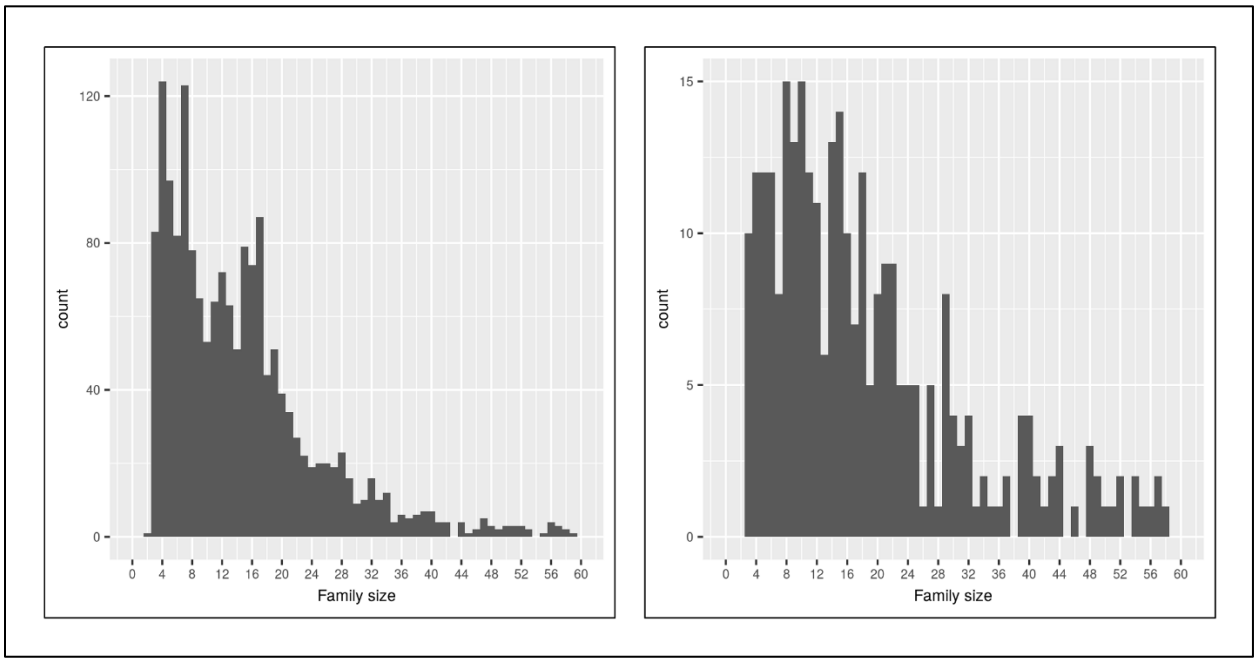

Fig 3.2. Family size distributions (up to size 60) of legume families involved in family merging (left) accepting previous unclustered sequences (right).

families, one between sizes 2 to 6 and another between sizes 16 to 18. This can be interpreted as small families with sizes 2 to 6 merging into families containing 16 to 18 sequences. We also see involvement of a considerable number of larger families in merging. Similarly, the size

distribution of unclustered corrected families also shows peaks between sizes 8 to 12, which shows the high tendency of these families in accepting previously unclustered sequences.

Next, the tree-based splitting procedure was applied on both merged and unmerged legume families to correct for over-clustering. For each family, closest outgroup sequences were first identified and Maximum-Likelihood (ML) family phylogeny was inferred along with the selected outgroup sequences. Each family tree was scored and analyzed for the presence of monophyletic legume clades after rooting the tree using the closest available outgroup sequence. Rooted trees with scores < 1.0 were analyzed for the presence of monophyletic legume clades. Families containing more than one legume clade were split in order to separate the clades into different families. As a result, 2554 merged and unmerged families were split into 5495 families. The size distribution, up to size 200, of the families that were split is shown in Fig 3.3. The distribution shows the majority of families that were split contained around 30 to 50 sequences according to the parameters used for splitting the families (See Methods).

We used the domain-composition-based family scoring to check for the improvements in the families produced after the application of the merging and splitting procedures. Since the domain-composition-based scoring is expected to assign higher scores to families with the constituent sequences containing similar domain compositions, the newly created families are expected to have, on average, higher domain composition scores than the original families. Therefore, both the original families and the corrected families were analyzed through the domain-composition-based scoring method to study the change in the domain compositions of the two family sets. Table 3.2 shows the increase in family counts in the new families, as compared to the original families, for domain composition scores ≥ 0.7. We can see an increase

in the number of families in the new set as compared to the original set, in all 4 high scoring categories.

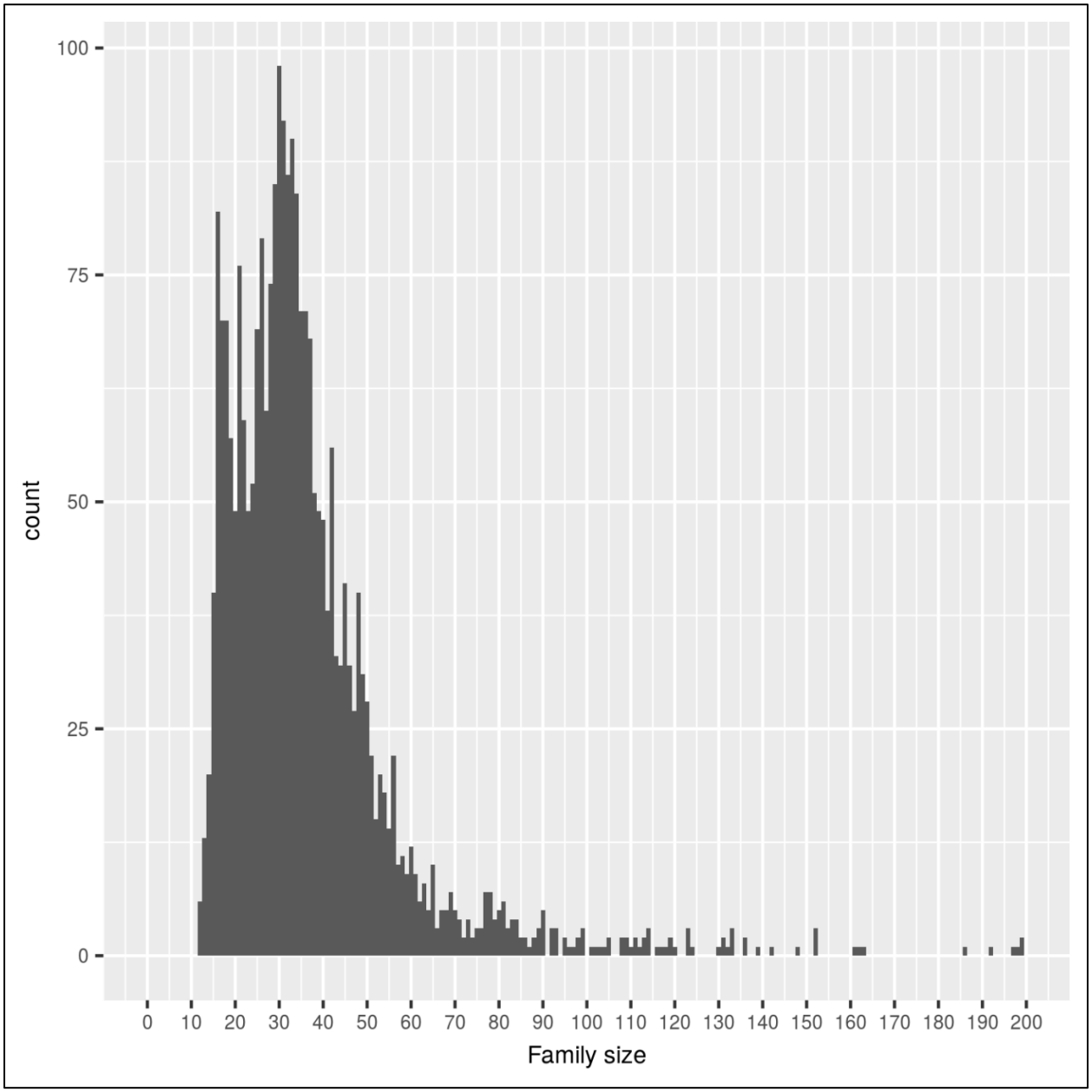

Fig 3.3. Family size distribution of legumes families that were split using the tree-based over-clustering correction method.

Table 3.2: Family counts in the new and the original set of legume families for high values of domain composition scores

score categories	# of new families	# of original families
= 1.0	6425	5542
≥ 0.9	12777	11328
≥ 0.8	14940	13226
≥ 0.7	15787	13997

Discussion

The legume gene families at legumeinfo.org [6] are built from 14 legume species using a custom family construction method that leverages information from the WGD at the base of the papilionoid subfamily. The family construction method uses the differences in synonymous sites (K_s) to detect and circumscribe families that diverged specifically due to occurrence of WGD in the papilionoid ancestor. In this work, we describe methods for improving gene families, testing these methods on the legume gene families. Our approach uses the following three steps: 1) merging over-fragmented families into larger families using a two-way HMM-based searching method, 2) placing previously unclustered sequences into families using the same HMM-based searching method, 3) splitting over-clustered families into separate clusters using a tree-based family scoring method. The application of the first and second steps resulted in merging of 1720 families into 841 clusters, and reclassification of 3,045 previously unclustered sequences into 347 families. Size distributions of the merged and the unclustered, corrected families showed the inclusion of many small families into larger families.

Subsequently, application of the third step to the merged and unmerged families resulted in separation of 2,554 families into 5,495 individual clusters. The family size distributions show

that majority of the families that were split contained 30 to 50 sequences - which is an expected size range for these families, considering speciation and genome duplication histories in the legumes. We also studied the improvements in the legume families obtained after the application of merging and splitting procedures using the protein domain composition scores of families. An increase in the number of high scoring families was observed in the new set, as compared to the original set which showed that families from new set were better in terms of domain compositions within the family. We release the new set families as an improved version of K_s-based legume families at this location: https://de.cyverse.org/dl/d/877F3083-0E4C-4A70-8624-E3AB14B3AA60/lgf5v2.tar.gz. Also, since the family merging and splitting techniques explained in this work operate directly on family clusters irrespective of the method used to produce the families, we also release the containerized versions of these techniques which can be downloaded from https://hub.docker.com/repository/docker/akshayayadav/hmmsearch-hmmscan-family-merging and https://hub.docker.com/repository/docker/akshayayadav/overcl-detection-correction, respectively. The docker container for scoring families using their protein domain compositions can be obtained from https://hub.docker.com/repository/docker/akshayayadav/genefamily-domain-composition-cosine-scoring.

References

1. Lewis GP (2005) Legumes of the World. Royal Botanic Gardens Kew

2. Azani N, Babineau M, Bailey CD, et al (2017) A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). TAXON 66:44–77

3. Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. Syst Biol 54:575–594

4. Bruneau A, Mercure M, Lewis GP, Herendeen PS (2008) Phylogenetic patterns and diversification in the caesalpinioid legumesThis paper is one of a selection of papers published in the Special Issue on Systematics Research. Botany 86:697–718

5. Cannon SB, McKain MR, Harkess A, et al (2015) Multiple Polyploidy Events in the Early Radiation of Nodulating and Nonnodulating Legumes. Mol Biol Evol 32:193–210

6. Stai JS, Yadav A, Sinou C, Bruneau A, Doyle JJ, Fernández-Baca D, Cannon SB (2019) *Cercis*: A Non-polyploid Genomic Relic Within the Generally Polyploid Legume Family. Front Plant Sci. https://doi.org/10.3389/fpls.2019.00345

7. Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, Liu X, Gao D, Clevenger J, Dash S (2015) The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat Genet 47:438

8. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83

9. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240

10. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J (2010) Genome sequence of the palaeopolyploid soybean. nature 463:178

11. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227–239

12. Hane JK, Ming Y, Kamphuis LG, et al (2017) A comprehensive draft genome sequence for lupin (Lupinus angustifolius), an emerging health food: insights into plant–microbe interactions and legume evolution. Plant Biotechnol J 15:318–330

13. Tang H, Krishnakumar V, Bidwell S, et al (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics 15:312

14. Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

15. De Vega JJ, Ayling S, Hegarty M, et al (2015) Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement. Sci Rep 5:17394

16. Kang YJ, Satyawan D, Shim S, et al (2015) Draft genome sequence of adzuki bean, *Vigna angularis*. Sci Rep 5:8069

17. Kang YJ, Kim SK, Kim MY, et al (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. Nat Commun 5:5443

18. *Vigna unguiculata* v1.1 (Cowpea). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Vunguiculata_er. Accessed 12 Feb 2019

19. Björklund ÅK, Ekman D, Light S, Frey-Skött J, Elofsson A (2005) Domain Rearrangements in Protein Evolution. J Mol Biol 353:911–923

20. Song N, Sedgewick R d., Durand D (2007) Domain Architecture Comparison for Multidomain Homology Identification. J Comput Biol 14:496–516

21. Bitard-Feildel T, Kemena C, Greenwood JM, Bornberg-Bauer E (2015) Domain similarity based orthology detection. BMC Bioinformatics 16:154

22. Eddy S (2003) HMMER User's Guide. Biological Sequence Analysis Using Profile Hidden Markov Models.

23. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780

24. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5:e9490

25. Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11:24

26. Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33:1635–1638

27. Finn RD, Coggill P, Eberhardt RY, et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285

28. Mistry J, Bateman A, Finn RD (2007) Predicting active site residue annotations in the Pfam database. BMC Bioinformatics 8:29

CHAPTER 4. *CERCIS*: A NON-POLYPLOID GENOMIC RELIC WITHIN THE GENERALLY POLYPLOID LEGUME FAMILY

Akshay Yadav, Jacob S. Stai, Carole Sinou, Anne Bruneau, Jeff J. Doyle, David Fernández-Baca, Steven B. Cannon

Modified from a manuscript published in *frontiers in Plant Science*

Abstract

Based on evolutionary, phylogenomic, and synteny analyses of genome sequences for more than a dozen diverse legume species as well as analysis of chromosome counts across the legume family, we conclude that the genus *Cercis* provides a plausible model for an early evolutionary form of the legume genome. The small *Cercis* genus is in the earliest-diverging clade in the earliest-diverging legume subfamily (Cercidoideae). The *Cercis* genome is physically small, and has accumulated mutations at an unusually slow rate compared to other legumes. Chromosome counts across 477 legume genera, combined with phylogenetic reconstructions and histories of whole-genome duplications, suggest that the legume progenitor had 7 chromosomes – as does *Cercis*. We propose a model in which a legume progenitor, with 7 chromosomes, diversified into species that would become the Cercidoideae and the remaining legume subfamilies; then speciation in the Cercidoideae gave rise to the progenitor of the *Cercis* genus. There is evidence for a genome duplication in the remaining Cercidoideae, which is likely due to allotetraploidy involving hybridization between a *Cercis* progenitor and a second diploid species that existed at the time of the polyploidy event. Outside the Cercidoideae, a set of probably independent whole-genome duplications gave rise to the five other legume subfamilies, at least four of which have predominant counts of 12–14 chromosomes among their early-diverging taxa. An earlier study concluded that independent duplications occurred in the

Caesalpinioideae, Detarioideae, and Papilionoideae. We conclude that *Cercis* may be unique among legumes in lacking evidence of polyploidy, a process that has shaped the genomes of all other legumes thus far investigated.

Introduction

The legume family, Leguminosae, with approximately 20,000 species, is the third most diverse plant family, after Orchidaceae and Asteraceae [1]. The family underwent a rapid radiation shortly after its origin ~59–64 million years ago (Mya) [2, 3], giving rise to six lineages that have recently been recognized as subfamilies by the international legume systematics community [1]. Among those subfamilies, four of them (Papilionoideae, Caesalpinioideae, Detarioideae, Cercidoideae) contain the vast majority of genera and species, while Dialioideae contains 17 genera and 84 species, and Duparquetioideae contains a single genus and species. The four larger subfamilies have been shown [4] to each have been affected by early whole-genome duplications (WGDs): at the base of the Papilionoideae and near the origins of the Cercidoideae, Detarioideae, and Caesalpinioideae – though the precise timing of the WGD(s) in the latter three lineages remains uncertain due to low sampling.

In particular, the WGD status and timing within the Cercidoideae has been uncertain: did a WGD predate the earliest divergences in the family, or did it occur later? Cannon et al. (2015) [4] reported a WGD signal for *Bauhinia tomentosa*, based on comparisons of divergence times of duplicated genes and orthologs based on synonymous substitution distributions (K_s peaks for duplication and speciation) from transcriptome sequence – but no WGD peak was evident for *Cercis canadensis*. This result was inconclusive, however: lack of a WGD peak could have been due to sequence loss or non-recovery for that genus. The genus *Cercis* is sister to the remainder

of the Cercidoideae genera [5–7]; we therefore address the question of whether *Cercis* was affected by an early WGD or whether the WGD occurred later in the evolution of the subfamily.

The legumes fall within the Fabidae (rosid 1) clade [8], and thus were affected by the gamma triplication event that occurred around the time of the origin of the core eudicots, approximately 120 Mya [9]. Species such as *Phaseolus* (bean; papilionoid) or *Desmanthus* (bundleflower; caesalpinioid) show evidence of old but independent duplications within the legume family [4]. Finding one or more early-diverging legume species without WGD would be of interest because such species could provide important clues to both the structure of the ancestral legume genome and the evolution of species and genomes across this large family.

In the present study, we investigate a new set of genome sequences from the Cercidoideae, Caesalpinioideae, and Papilionoideae, as well as extensive chromosome count data from across the legumes. We also describe results from targeted sequencing of selected genes within the Cercidoideae, to clarify the timing and nature of WGDs affecting the legumes. We present evidence supporting lack of a WGD in the genus *Cercis*, and hypothesize an allotetraploidy event affecting the remainder of the Cercidoideae subfamily.

Materials and Methods

Gene Family Construction, K_s Analysis, and Phylogeny Calculation

Gene families include proteomes (complete sets of translated coding sequences – one representative transcript per gene) from fifteen legume species, and five non-legume species – which were used for phylogenetic rooting and evolutionary context. Species and sources are indicated in Table 4.1. We used a custom gene family construction method in order to best capture some challenging features of the phylogeny. Gene family features to account for include early WGDs affecting species in the family – but we wished to avoid an older genome

triplication, occurring early in angiosperm evolution. Therefore, we used a combination of homology filtering based on per-species synonymous site changes, comparison with outgroup species, Markov clustering, and progressive refinements of family hidden Markov models (HMMs). The gene families are available at https://legumeinfo.org/data/public/Gene_families/legume.genefa m.fam1.M65K/ and associated methods and scripts are available at https://github.com/LegumeFederation/legfed_gene_families although the resources at those locations are focused on papilionoid species rather than on the non-papilionoid species examined in this paper. The same gene families above were used in the analysis in this paper, but with several papilionoid species removed and five other species added (via HMM-search and HMM alignment of the other species to the gene-family HMMs), as shown in Table 4.1 [10–24].

Table 4.1. Genome and annotation sources and versions.

Species	Genotype	Assembly	Annot.	Publication	Source
Arachis duranensis	V14167	1	1	Bertioli et al. (2016)	PeanutBase
Arachis ipaensis	K30076	1	1	Bertioli et al. (2016)	PeanutBase
Cajanus cajan	ICPL87119	1	1	Varshney et al. (2012)	LegumeInfo
Glycine max	Williams 82	2	1	Schmutz et al. (2010)	Phytozome
Phaseolus vulgaris	G19833	2	1	Schmutz et al. (2014)	Phytozome
Vigna radiata	VC1973A	6	1	Kang et al. (2014)	LegumeInfo
Lotus japonicus	MG20	3	1	Sato et al. (2008)	Phytozome
Medicago truncatula	A17_HM341	4	2	Tang et al. (2014)	Phytozome
Cicer arietinum	Frontier	1	1	Varshney et al. (2013)	LegumeInfo
Nissolia schottii		1	1	Griesmann et al. (2018)	GigaDB
Mimosa pudica		1	1	Griesmann et al. (2018)	GigaDB

Table 4.1 Continued					
Species	**Genotype**	**Assembly**	**Annot.**	**Publication**	**Source**
Chamaecrista fasciculata		1	1	Griesmann et al. (2018)	GigaDB
Bauhinia tomentosa		1	1	Cannon et al. (2015)	GigaDB
Cercis canadensis		1	1	Griesmann et al. (2018)	GigaDB
Prunus persica	Lovell	2	2.1	IPGI (2013)	Phytozome
Cucumis sativus		1	1	Phytozome, 2017	Phytozome
Vitis vinifera	PN40024	12X	12X	Jaillon et al. (2007)	Phytozome
Arabidopsis thaliana	Col-0	TAIR10	TAIR10	Berardini et al. (2015)	Phytozome
Solanum lycopersicum	LA1589	ITAG2.4	ITAG2.4	Tom. Gen. Cons. (2012)	Phytozome

Gene families were generated as follows. All-by-all comparisons of protein sequences for all species were calculated using BLAST [25]). Matches were filtered to the top two matches per query, with at least 50% query coverage and 60% identity. For the resulting gene pairs, in-frame nucleotide alignments of coding sequences were calculated, which were used, in turn, to calculate synonymous K_s counts per gene pair, using the PAML package [26], with the Nei and Gojobori (1986) [27] method for estimating the numbers of synonymous nucleotide substitutions. The calculation process was driven using the synonymous_calc.py wrapper script [28], which additionally uses the packages biopython [29], ClustalW2 [30], and PAL2NAL [31]. For each species pair, histograms of K_s frequencies were used as the basis for choosing per-species K_s cutoffs for that species pair in the legumes. For most species pairs, the selected peak corresponded with the papilionoid duplication (K_s average of 0.6, varying between 0.45 and 0.8). For comparisons between papilionoid species and the four non-papilionoid legume species (*Mimosa pudica*, *Chamaecrista fasciculata*, *B. tomentosa*, and *C. canadensis*), the selected peak

corresponded to the speciation divergence between the pair of species. To accommodate variation in K_s values, the cutoff for each species pair was generally set at 1.5 times the modal K_s value (K_s peak). The set of gene pairs was filtered to remove all pairs with K_s values greater than the per-species-pair K_s cutoff. The resulting set of filtered pairs was used for Markov clustering, implemented in the mcl program [32], with inflation parameter 1.2, and relative score values (transformed from K_s values) indicated with the -abc flag. Sequence alignments were then generated for all gene families using MUSCLE [33]. Hidden Markov models (HMMs) were calculated from the alignments using the hmmer package [34], and sequences in each family were realigned to the family that those sequences were assigned to, in order to determine HMM bitscores and calculate a median alignment score for each family. Families were then evaluated for outliers: sequences scoring less than 40% of the median HMM bitscore for the family were removed. The HMMs were then recalculated for each family (without the low-scoring outliers), and were used as targets for HMM search of all sequences in the proteome sets – including those omitted during the initial K_s filtering. Again, sequences scoring less than 40% of the median HMM bitscore for the family were removed. These HMM alignments were then used for calculating phylogenetic trees, after trimming non-aligning characters (characters outside the HMM match states). Phylogenies were calculated using RAxML [35], with model PROTGAMMAAUTO, and rooted using the closest available outgroup species.

Calculation of K_s Values and Modal K_s Peaks

Synonymous-site differences (K_s) were calculated by two methods: first, based on gene-pairs derived from the top two matches of genes between or within species, based on blastp sequence searches; and second, based on gene-pairs derived from genomic synteny comparisons and coding-sequence coordinates, provided to the CoGe SynMap service at

https://genomevolution.org/coge/ [36]. In the former case (calculated on top blastp matches), K_s values were calculated using PAML, driven by synonymous_calc.py, by Haibao Tang, available at https://github.com/tanghaibao/bio-pipeline. From the PAML output, the Nei-Gojobori K_s value was used [27]. For both approaches (BLAST-based and synteny gene-pair-based), K_s histograms were calculated after filtering for K_s values between 0 and 2.

Inference of Consensus Branch Lengths from K_s Peaks

To infer branch lengths for an idealized gene tree from these K_s peak values (Fig 4.1D), modal K_s peak values were read from K_s histograms, with values representing WGD events for a species compared with itself (e.g., in *Phaseolus* with respect to the papilionoid WGD) or orthologous gene separations between species (e.g., between *Phaseolus* and *Cercis*). The modal K_s values were then used to algebraically calculate branch lengths along a gene tree with known species topology and hypothesized duplication history, for the selected species. In these calculations, each branch segment is a variable to be solved, given the observed distances between each terminal (e.g., 0.55 for the phylogenetic path between *Phaseolus* and *Cercis*). Because the internal branch lengths are not uniquely determinable from the observed K_s path-lengths, several branch lengths were set at 0.01 (based on very short branch lengths observed in both gene trees and species trees): branches subtending the *Chamaecrista* WGD, the papilionoid/caesalpinioid clade, and the *Cercis*–*Bauhinia* 2 clade. Then, a PHYLIP-format [37] gene tree was manually generated for the represented species, using branch length values from the algebraic calculations.

Methods for Mining for Tree Topologies

To test the order of phylogenetic events, gene trees were evaluated for 14,709 legume gene family trees that contain *Cercis* and/or *Bauhinia* sequences. Python scripts that use the functions from the ETE Toolkit [38, 39] were used to read and analyze the legume gene family trees using the species overlap method [40]. The species overlap method labels an internal node in a given rooted tree as D (duplication event) or S (speciation event) based on whether there are common species between both partitions corresponding to the two subsequent children nodes. Species-overlap tests were run for trees in which same-species terminal pairs were collapsed (when both branch lengths were less than 0.01), to control for local private gene duplications.

Results

K_s Peaks from Self-Comparisons of Coding Sequence

Within- and between-species comparisons of rates of synonymous-site changes per synonymous site were evaluated by Cannon et al. (2015) [4] for 20 diverse legume species – including representatives from each of the four largest legume subfamilies. These showed K_s peaks of around 0.3–0.6 in all species except *Cercis*, where only a much older peak of ~1.5 was seen. Because that work was based on transcriptome sequence for most species, there was some question whether the absence of the peak in *Cercis* might be due to poor sequence quality or sequence non-recovery (although the transcriptome assembly statistics were generally in the same range as for the other species). Recent availability of genome sequences for *C. canadensis*, *C. fasciculata*, *M. pudica*, and *Nissolia schottii*, from Griesmann et al. (2018) [18], provides an opportunity to test K_s and other results with greater rigor. *Chamaecrista* and *Mimosa* fall within the Caesalpinioideae subfamily, and *Nissolia* is in the Papilionoideae subfamily, within the dalbergioid clade, along with peanut (*Arachis*). For K_s analysis in this study, we focus

particularly on *Cercis*, *Bauhinia* (as representatives of the Cercidoideae), *Chamaecrista* (as a representative from the Caesalpinioideae), and *Phaseolus* (as a representative of the Papilionoideae), to investigate evidence for the presence and timing of possible WGDs in these lineages. We include *Phaseolus* to provide an example of a species with high-quality genome sequence and a well-studied, early WGD.

K_s results from genes predicted in the *C. canadensis* ("cerca") and *C. fasciculata* ("chafa") genome assemblies are shown in Fig 4.1, along with genes from *Phaseolus vulgaris* ("phavu") and from *B. tomentosa* ("bauto"; transcriptome-derived). The K_s values were determined both for top BLAST-based gene-pairs between species and within species (e.g., top pairs within *Cercis*).

There is a clear K_s peak for *Cercis–Bauhinia* at 0.15 and a peak for *Bauhinia* compared with itself at 0.25 (Fig 4.1A, 4.1C). Although there are some duplications near 0 in *Cercis* compared with itself, there is no older *Cercis–Cercis* peak as the prominent peak seen in *Bauhinia–Bauhinia* at 0.25. The duplications near 0 in the *Cercis–Cercis* plot are likely due to local gene duplications (as also seen, for example, in the *Phaseolus–Phaseolus* self-comparison in Fig 4.1A vs 4.1B), as this signature of recent duplications is absent in the synteny-derived K_s plots in Fig 4.2.

We find the expected strong WGD peak within *Phaseolus* and also for *Phaseolus–Cercis* (at 0.6 and 0.55), respectively, but again, no older peak within *Cercis* compared with itself (Fig 4.1B). The fact that the *Phaseolus–Phaseolus* modal K_s peak is greater than the *Phaseolus–Cercis* peak suggests a much greater rate of mutation accumulation in *Phaseolus* and its progenitors in Papilionoideae than in *Cercis* and its progenitors in Cercidoideae [41, 42].

In Fig 4.1C, there is a speciation peak for *Phaseolus–Bauhinia* that is similar to *Phaseolus–Cercis* with the exception that the *Phaseolus–Bauhinia* peak appears slightly "older" than for *Phaseolus–Cercis* (0.6 vs. 0.55), suggesting more rapid rate of mutation accumulation in *Bauhinia* than in *Cercis*. Fig 4.1D shows an inferred consensus gene tree, with branch lengths calculated (with approximation) from K_s plots in Fig 4.1, 4.2 (as described in Methods).

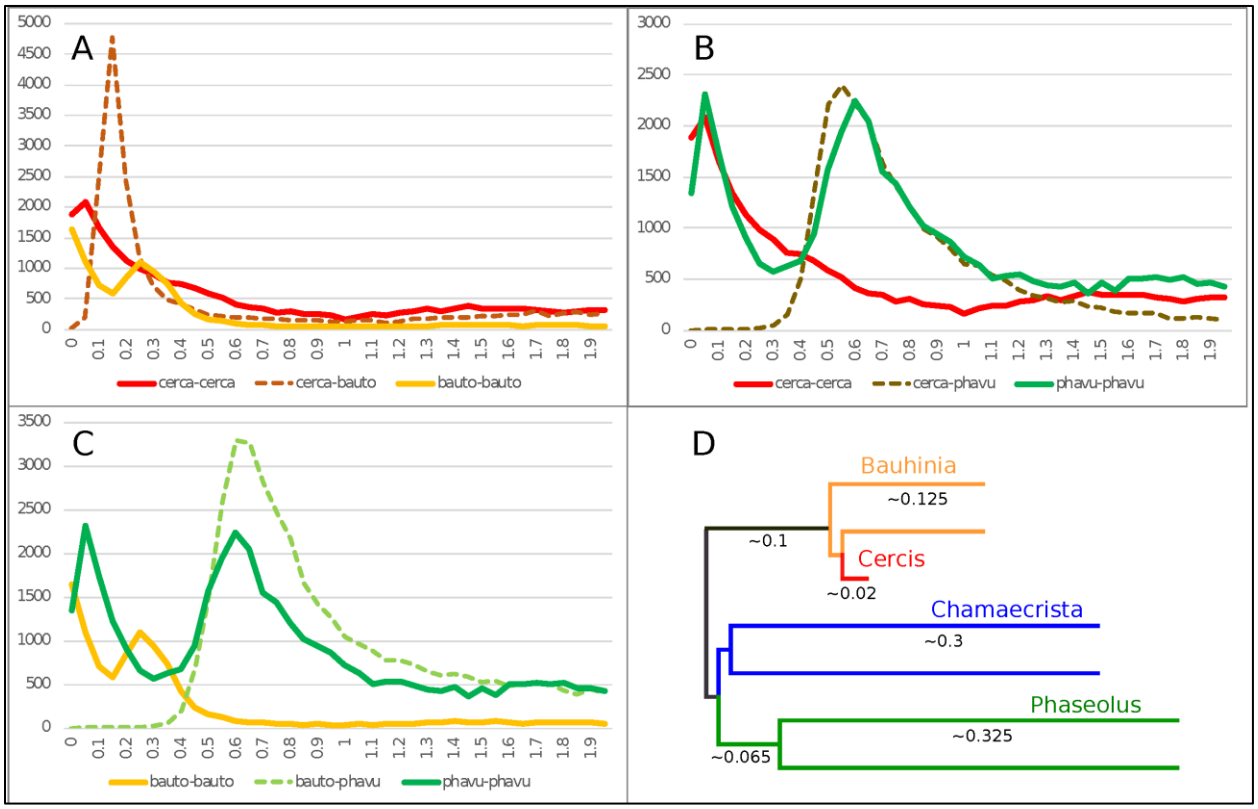

Fig 4.1. Histograms of K_s values for top gene-pair comparisons for *Cercis canadensis* ("cerca"), *Bauhinia tomentosa* ("bauto"), and *Phaseolus vulgaris* ("phavu"). In K_s plots **(A–C)**, solid lines are for self-comparisons (e.g., for *Cercis* gene-pairs), and dotted lines are for between-species comparisons (e.g., between *Cercis* and *Bauhinia*). The schematic tree in panel **D** is an idealized distance tree in which each OTU represents an "average" gene: either a single copy in *Cercis*, or each of two homoeologs created by unique WGD events in the remaining taxa. Branch lengths are calculated from pairwise modal K_s values in panels **A–C**.

In Fig 4.2A–C, K_s values are derived from gene-pairs within synteny blocks derived from genome comparisons. A major effect of this strategy is to exclude local gene duplications – and to reduce other paralogous matches that can show up as recent duplications – for example, in matches among many members of a recently expanded gene family. This reduction in recent- and locally derived paralogs is evident in K_s counts near zero for "young" (small) K_s values. The sloping K_s histogram seen in Fig 4.1 for *Cercis–Cercis* is entirely absent in Fig 4.2. The modal K_s "peak" for *Cercis*, if there is any, is in the range of 1.5–2 – contrasting with the *Cercis–Phaseolus*, *Cercis–Chamaecrista*, and *Chamaecrista–Phaseolus* peaks of 0.6, 0.5, and 0.7, respectively – indicating that any *Cercis* WGD peak in this data would well predate the legume origin.

Also noteworthy in Fig 4.2 is the low modal K_s peak for *Chamaecrista–Chamaecrista* (amplitude of 101, compared with 581 for *Phaseolus–Phaseolus*). This difference in numbers of paralogous duplicated genes could be due to higher rates of gene loss from *Chamaecrista* following WGD early in the Caesalpinioideae. The strong K_s peaks in the orthologous *Chamaecrista – Cercis* comparison and the *Phaseolus – Cercis* comparison suggest that there is nothing systematically wrong with the *Chamaecrista* gene models. Rather, it appears that *Chamaecrista* is more fully "diploidized," with a higher proportion of duplicated genes having reduced to single copies, providing a sufficient basis for discovering correspondences with other species, but erasing much of the WGD signature in a *Chamaecrista* self-comparison. Similar diploidization and interspersed gene losses have been reported in *Medicago truncatula* [43].

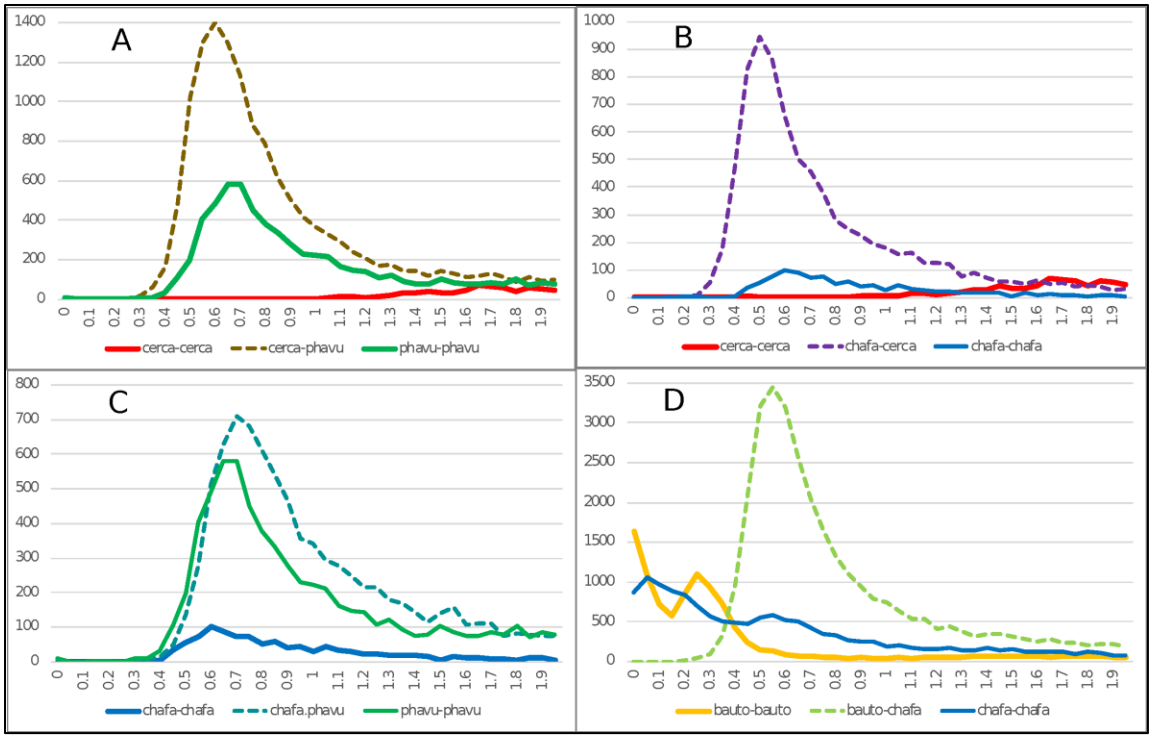

Fig 4.2. Histograms of K_s values for synteny-based comparisons for *C. canadensis* ("cerca"), *Chamaecrista fasciculata* ("chafa"), *P. vulgaris* ("phavu"), and *B. tomentosa* ("bauto"). In K_s plots, solid lines are for self-comparisons (e.g., for *Cercis* gene-pairs), and dotted lines are for between-species comparisons (e.g., between *Cercis* and *Phaseolus*). This Fig differs from Fig 4.1 both in species selection and in method for selecting gene pairs: in Fig 4.1, K_s values are calculated for all top gene pairs, and in panels **A–C**, K_s values are calculated for gene-pairs from synteny features identified from genomic comparisons (panel **D** is an exception: the K_s values are calculated from all top gene pairs, because only transcriptomic sequence is available for *Bauhinia*). The effect of using synteny-based gene pairs for calculating K_s is apparent in the *Chamaecrista* self-comparison plots (chafa–chafa; blue) in panel **B** or **C** (syntenic-based) vs. panel **D** (gene-pair based): in the gene-pair based figures in D, the WGD peak is still evident at ~0.55–0.6, but the signal from more recent gene pairs are also apparent – presumably, as a result of independent, local gene duplications within *Chamaecrista*.

Genomic Synteny Analysis

Given the draft genomic sequence assembly for *Cercis*, it is possible to make synteny comparisons with other legume genome assemblies, as well as assemblies of near outgroups to the legumes. In a synteny comparison of two genomes, a WGD present in one of the genomes and absent in the other should be apparent in a genomic dotplot through the following pattern: starting from a given genomic region in the non-duplicated genome and tracing through the dotplot, one should find matches to two regions in the genome with the WGD; and starting from a given genomic region in the duplicated genome and tracing through the dotplot in the other axis, one should find matches to a single region in the genome that lacks the WGD. This can be described in terms of "synteny depth:" the depth of the duplicated genome should be twice that of the non-duplicate genome.

Because the *Cercis* assembly is still highly fragmented (N50 of 421 kb), synteny depth is difficult to assess visually, but it can be measured computationally. The quota-alignment package [44] identifies synteny blocks between two genomes, attempting to match a specified pair of synteny depths or "quotas.". For example, if genome B has a WGD that A lacks, then the quota for B relative to A would be 2:1. If the quota is mis-specified as 1:1, then a poor coverage score will result for the duplicated genome, because many potential blocks in genome B will be missed. We also note that in the quota-alignment package, in a genome self-comparison, the trivial self-match is suppressed, so the expected quota for a genome with a single WGD, compared with itself, would be 1:1 rather than 2:2.

We used the quota-alignment package to test a range of quotas for all comparisons among *Cercis*, *Phaseolus*, and *Prunus*. There is no evidence for a duplication in *Prunus* since the angiosperm whole-genome triplication (WGT) [9, 20], and there is a known WGD in *Phaseolus* at around 50 Mya [4, 42], so these should serve as useful comparisons relative to *Cercis*. For

Prunus–Phaseolus, a quota of 1:1 gives *Phaseolus* coverage of only 63.8% (Table 4.2) vs. 96% for *Prunus*, indicating that less than two-thirds of the *Phaseolus* genome has synteny coverage for the identified gene pairs. A quota of 1:2 for *Prunus–Phaseolus* is much better, at 97.4 and 96.8% coverage, respectively. For *Prunus–Cercis*, a quota of 1:1 gives acceptable coverage of 93.4 and 95.2%, respectively; a quota of 1:2 improves the coverage by only about 2% (Table 4.2). For *Phaseolus–Cercis*, the best quota is 2:1, with coverages of 93.3 and 94.7%, respectively. For the self-comparisons for each species, there is notable improvement going from 1:1 to 2:2 (Table 4.2). This is likely due to the ancient angiosperm triploidization [9], which generated three genome copies; the expected number of synteny blocks from any region would then be two (ignoring the trivial self-match).

The K_s peak values derived from gene pairs in the synteny analysis (Table 4.2) are consistent with the synteny depth results – with the *Cercis–Cercis* peak being of comparable age to *Prunus–Prunus* (1.74 and 1.4, respectively), and likely both dating to the angiosperm WGT. In contrast, the peak for *Phaseolus–Phaseolus* is 0.7, consistent with the papilionoid WGD. Taken together, the synteny and K_s results from Table 4.2 indicate that *Cercis* has the same overall WGD depth as *Prunus* and half that of *Phaseolus*, in comparisons among these genomes. In other words, the synteny and K_s evidence supports lack of a WGD in *Cercis*.

Phylogenomic Analyses

To determine duplication events in a phylogenetic context, we constructed gene trees for all legume genes, for fifteen diverse legume species: *Glycine max*, *P. vulgaris*, *Vigna unguiculata*, *Lupinus angularis*, *Arachis ipaensis*, *N. schottii*, *Cicer arietinum*, *M. truncatula*, *Lotus japonicus*, *C. fasciculata*, *M. pudica*, *B. tomentosa*, and *C. canadensis*. The first nine of these are from the Papilionoideae (representing the millettioid, genistoid, dalbergioid, and IRLC

Table 4.2. Synteny coverage for comparisons between the genomes of *Cercis canadensis*, *Phaseolus vulgaris*, and *Prunus persica*, at selected synteny "quotas" (expected coverage depths). For the comparison between *Prunus* and *Phaseolus* (with known WGD histories), the best quota choice is 1:2, corresponding with two synteny blocks in *Phaseolus* for one in Prunus. Similarly, for the comparison between *Cercis* and *Phaseolus*, the best quota choice is 1:2, corresponding with two synteny blocks in *Phaseolus* for one in *Cercis*; and for the comparison between *Cercis* and *Prunus*, the best quota choice is 1:1, suggesting that neither genome has a recent WGD in its history. The K_s peak values are consistent with this conclusion – with the *Cercis*-*Cercis* being of comparable age to *Prunus*-*Prunus* (and likely dating to the angiosperm whole-genome triplication).

Quotas	**X** Cercis	**Y** Cercis	***K*s peak**	**Comments**
q1-1	87.1	87.8	1.74	OK
q2-2	99.9	99.9		BEST
	Phaseolus	Cercis		
q1-1	61.9	94.1	0.62	At q1:1, Phaseolus coverage is too low
q2-1	93.3	94.7		BEST
	Prunus	Cercis		
q1-1	93.4	95.2	0.92	OK
q1-2	94.1	97.8		little improvement over q1:1
q2-2	99.2	98.6		BEST
	Phaseolus	Phaseolus		
q1-1	91.7	92.0	0.70	OK
q2-2	98.9	98.9		BEST
	Prunus	Phaseolus		
q1-1	96.0	63.8	1.16	At q1:1, Phaseolus coverage is too low
q1-2	97.4	96.8		BEST
	Prunus	Prunus		
q1-1	84.7	84.2	1.40	OK
q2-2	99.6	99.2		BEST

clades). We also included five non-legume outgroups – using one sequence from each, for each family, in order to provide a rooting for the legume sequences: *Arabidopsis thaliana*, *Prunus persica*, *Cucumis sativus*, *Solanum lycopersicum*, and *Vitis vinifera*. For convenience, analyses and figures that use sequences from these species use the following abbreviation form to indicate genus and species: the first three letters of the genus and the first two letters of the species epithet, e.g., "glyma" for G. max. Gene families were calculated to span the depth of the legume most-recent common ancestor – i.e., avoiding fragmented gene families that split sequences that have a common proto-legume ancestor, and avoiding over-clustered families that include legume sequences that diverged prior to the legume origin. Our method produced 18,543 such families, but for the present analysis, we analyzed the 14,709 families that contain one or more sequences from *Cercis* and/or *Bauhinia*. The set of 14,709 were used for subsequent phylogenomic analyses.

Informal Observations About Patterns in Trees

Gene family trees containing *Cercis* and *Bauhinia* sequences were used to investigate the occurrence of WGD in the most recent common ancestor (MRCA) of the *Cercis* and *Bauhinia* lineages. Although the phylogenomic analysis was likely complicated by uncertainties in phylogenetic reconstructions and by sequence losses or non-recovery, there are clear patterns in the results. We repeatedly see topologies congruent with those in two gene families shown in Fig 4.3 (families 31DXWY and 2SH9KY; names from this set of legume gene families were assigned random "license plate" names of six alphanumeric characters). These gene families each show two *Bauhinia* sequences and one *Cercis* sequence in one clade. Both gene families show duplicated sequences for *Mimosa* and *Chamaecrista* (Caesalpinioideae; although in Fig 4.3A, these do not resolve to a single clade, which may indicate that the duplication occurred

very early in the Caesalpinioideae) in the Papilionoideae, there are paired sequences from most species, highlighting the pre-papilionoid WGD [4]. In the Cercidoideae clade, there is a curious feature: the duplication that affects *Bauhinia* predates the *Bauhinia–Cercis* speciation, and produces the expected two homoeologs in *Bauhinia*, but there is only a single *Cercis* sequence.

Summaries of Sequence Counts for All Gene Families (Legume Phylogeny Working Group et al., 2017)

To investigate WGDs in the legumes, we analyzed gene counts across all legume gene families. A summary overview of the phylogenomic analysis is shown in Table 4.3, which gives counts of gene families (and trees) having the indicated sequence count for each species (Only selected species are shown in Table 4.3). These are given for two variants of the trees: first (A) for the full, unmodified trees, and second (B) for trees in which similar ($K_s < 0.2$) terminal sequence pairs for a species have been reduced to a single representative, in order to reduce the effect of private, genus-specific WGDs. For example, in Table 4.3A, the first column (glyma / G. max) shows the largest number of trees (6531) having two sequences, and the second largest number of trees (3995) having four or more sequences. A count of four for G. max would be expected in a gene family in which no gene loss occurred following the two WGDs in the *Glycine* lineage within the period of legume evolution [45]. In Table 4.3B, in which terminal same-species pairs have been reduced to a single representative, the largest number of trees (7951) has one sequence, and the second largest number of trees (4217) has two sequences.

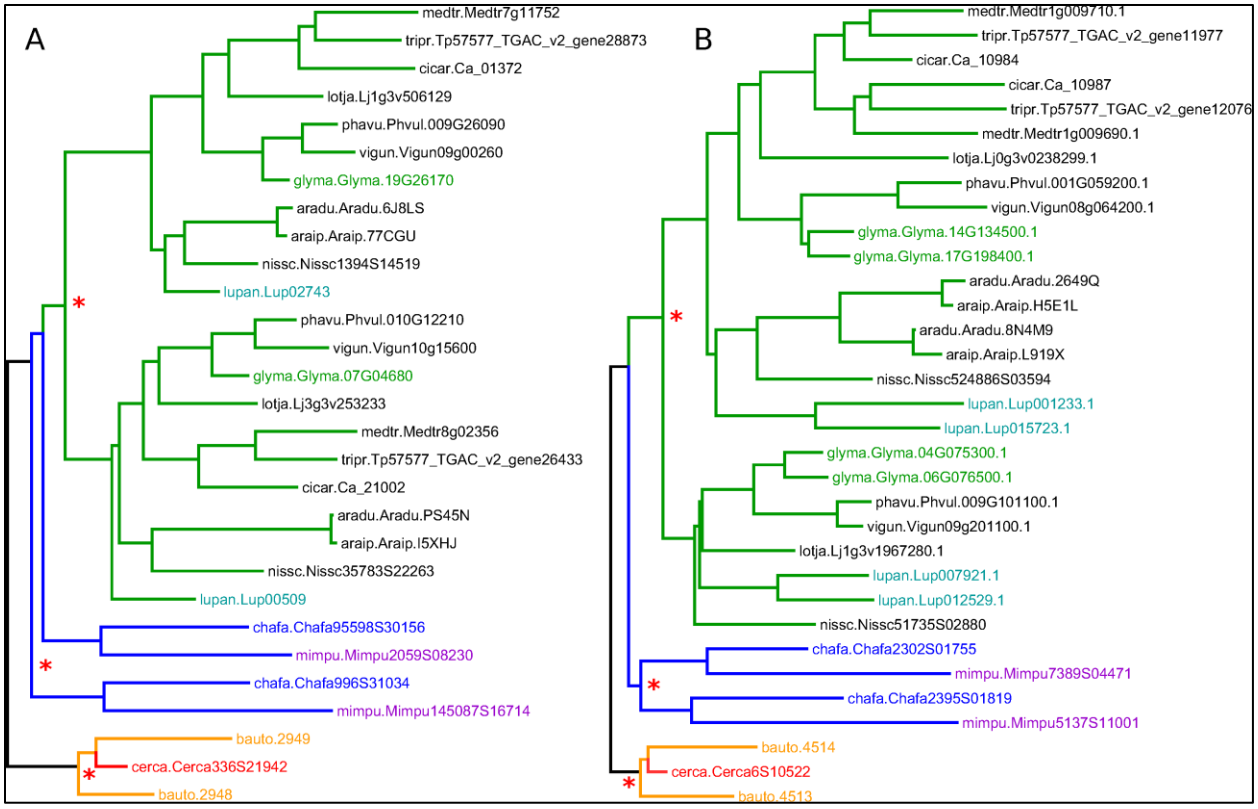

Fig 4.3. Sample gene trees (for gene families 31DXWY and 2SH9KY; A and B, respectively), showing clades corresponding to the Cercidoideae (orange and red), Caesalpinioideae (blue and violet), and Papilionoideae (green). Species abbreviations are composed of the first three letters from the genus and the first two letters of the species. Full name correspondences are indicated in the text. Non-legume outgroup sequences are in gray. Red asterisks mark common ancestors of homoeologous sequence pairs. Additional, more recent WGDs within the Papilionoideae are highlighted with colors of the sequence IDs: green for *Glycine max* and turquoise for *Lupinus angustifolius*.

We propose that an indicator of potential older WGDs for a species is obtained by dividing the number of gene family counts for which a species is represented at least twice in the family by the number of family counts for which a species is represented only once. These ratios are given at the bottom of Tables 4.3A, 4.3B. For species with a WGD within the period of legume evolution, a relatively larger number of families should have two or more sequences. The most dramatic ratio is for *Glycine* (632%; i.e., 6.3 × the naïve expectation) – which has two WGDs in its legume history (pre-papilionoid and a much more recent Glycine-specific duplication). For the unreduced trees (1A), all other species have ratios greater than 50% except for *Cercis*, with 24%. For the reduced trees (with collapsed terminal same-species clades), the ratios are somewhat lower for all species: 42–78% for all species except *Cercis*, with 20%. We interpret these results as evidence for WGD in all of the represented legume species except *Cercis*.

Table 4.3A. Counts for original full trees.

count	glyma	phavu	aradu	nissc	medtr	tripr	lotja	chafa	mimpu	bauto	cerca
0	553	826	2264	1425	1001	1252	1873	2558	3859	4066	1557
1	1933	8748	7761	8472	8141	8255	7602	7894	6432	5921	10567
2	6531	3981	3390	3656	3545	3429	3444	3178	2858	2570	1708
3	1697	716	752	681	984	957	1138	591	846	1130	437
≥4	3995	438	542	475	1038	816	652	488	714	1022	440
≥2 / =1	632%	59%	60%	57%	68%	63%	69%	54%	69%	80%	24%

Table 4.3B. Counts for trees with terminal recent pairs per species are reduced to a single representative.

count	glyma	phavu	aradu	nissc	medtr	tripr	lotja	chafa	mimpu	bauto	cerca
0	553	826	2265	1427	1003	1254	1873	2559	3860	4067	1558
1	7951	9034	7907	8815	8806	8878	9018	8353	7934	7475	10988
2	4217	3911	3396	3621	3443	3285	3066	2970	2160	2362	1564

Table 4.3B Continued											
count	glyma	phavu	aradu	nissc	medtr	tripr	lotja	chafa	mimpu	bauto	cerca
3	1163	616	707	545	798	791	534	484	430	546	342
≥4	825	322	434	301	659	501	218	343	325	259	257
≥2 / =1	78%	54%	57%	51%	56%	52%	42%	45%	37%	42%	20%

Mining for Tree Topologies Within the Cercidoideae

To infer the relative timing of gene duplications relative to speciations, we mined legume gene phylogenies for topological patterns expected to be produced by these events. Monophyletic groups were detected from a set of 14,709 families containing at least one sequence each from *Cercis* and *Bauhinia* (Fig 4.4 and Table 4.4). The MRCA node for each clade containing *Cercis* and *Bauhinia* was labeled either as D (for a duplication event) or S (for a speciation event), based on whether there are common species between both partitions corresponding to the two subsequent children nodes. For example, considering clades with two sequences from each of *Bauhinia* and *Cercis*, [(B, C), (B, C)] would be labeled D while [(B, B), (C, C)] would be labeled S (Fig 4.4) The species overlap method has been previously used to study evolutionary relationships of human proteins with their respective homologs in other eukaryotes [40]. We considered three types of monophyletic groups varying by number of *Cercis* and *Bauhinia* sequences: clades containing ≥ 2 *Cercis* and ≥ 2 *Bauhinia* sequences, clades containing exactly 1 *Cercis* and ≥ 2 *Bauhinia* sequences, and finally clades containing exactly 1 *Bauhinia* and ≥ 2 *Cercis* sequences. The proportions of clades out of the total number of clades, for all the three types, that were labeled as D at the MRCA node were also calculated. Species-overlap tests were run on trees in which very recently derived same-species terminal pairs were collapsed (when both branch lengths were less than 0.01), to control for local private gene duplications.

There are approximately tenfold more trees with one *Cercis* and two or more *Bauhinia* sequences than with one *Bauhinia* and two or more *Cercis* sequences (Table 4.4; 425/3205 and 183/2036). We interpret this result (preponderance of the 1 *Cercis*, ≥ 2 *Bauhinia* pattern) as evidence for WGD in *Bauhinia* but not *Cercis*. Further, of the clades with two or more *Bauhinia* sequences and one *Cercis* sequence, most (63%) of these have *Cercis* nested within the clade: 2036 of the total clade count look like [(B, C), B] rather than [(B, B), C] – the former likely resulting from a duplication of *Bauhinia* prior to speciation, and the latter resulting from speciation followed by duplication of *Bauhinia*. This result might seem nonsensical (duplication predating the *Cercis–Bauhinia* speciation, yet not affecting *Cercis*), but it would be consistent with allopolyploidy – with a *Cercis* progenitor having contributed one of the subgenomes in the allopolyploidy event that gave rise to *Bauhinia* and all other species in the rest of the Cercidoideae clade (elaborated further in the section "Discussion").

Gene Duplication Patterns Across Diverse Species in the Cercidoideae

To determine gene duplication patterns for species in the Cercidoideae, we take advantage of the well-conserved CYCLOIDEA-like TCP genes, which have been used both for phylogenetic inference and for studies of evolutionary development in the legumes [46, 47]. Using two sets of degenerate PCR primers that preferentially amplify two classes of CYCLOIDEA-like TCP genes in the legumes [46], Sinou and Bruneau (pers. comm.) amplified CYCLOIDEA-like genes from 114 species in Cercidoideae. These span all twelve genera in this subfamily. A phylogeny from a subset of these sequences is shown in Fig 4.5 – with sequences from each genus included but omitting some species from well-represented genera.

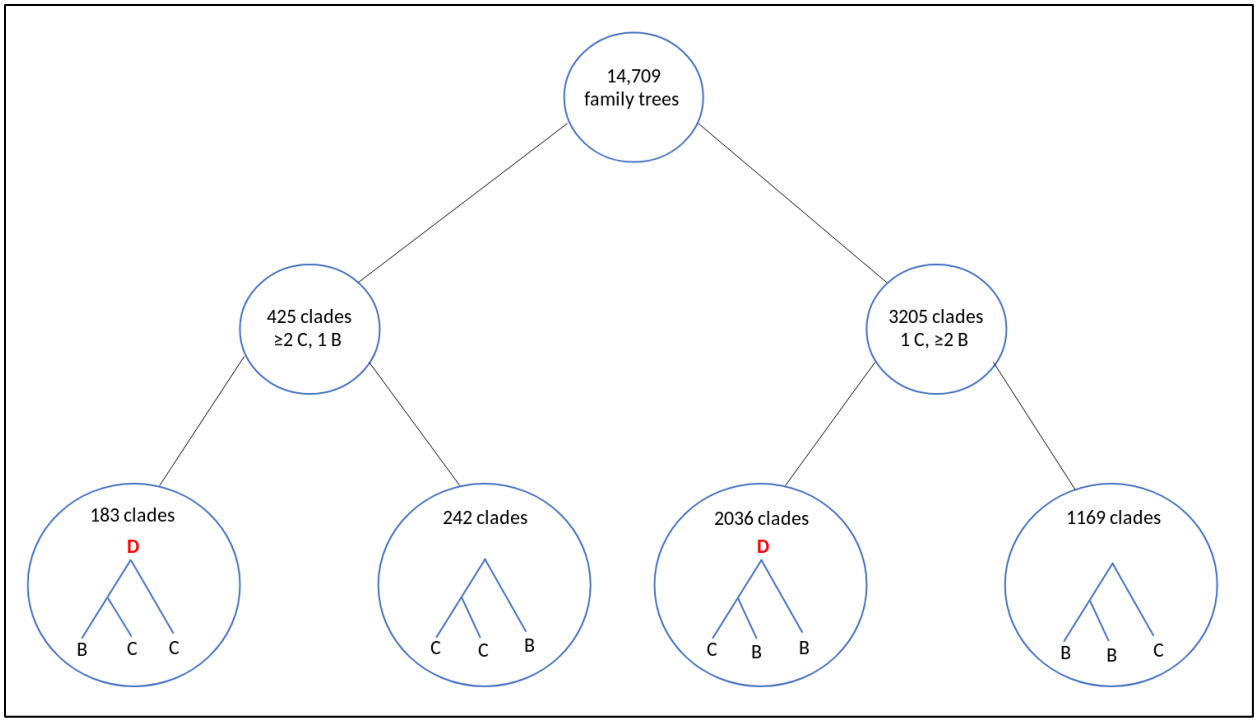

Fig 4.4. Graphical depiction of tree-mining results for topologies in the Cercidoideae. From 14,709 family trees with *Cercis* and *Bauhinia* sequences, clades with ≥2 *Cercis* and one *Bauhinia* sequence were 7.5 times more common than clades with 1 *Cercis* and ≥2 *Bauhinia* sequences (425 vs. 3205 clades, respectively). Of the latter (more frequent) clade configuration, cases with [(C, B), B] are 1.74 times more common than cases with [(B, B), C] (2036 vs. 1169 clades, respectively). In the first of these patterns, [(C, B), B], the MRCA node of the clade is labeled as a Duplication by the "species overlap" algorithm (see section "Materials and Methods" for description) – meaning that a the MRCA is inferred as due to a gene duplication event rather than a speciation-derived orthology event. Asterisks mark nodes where orthologous genes derive from speciation. Also see Table 4.4 for counts and percentages.

Table 4.4. The types of monophyletic groups containing different numbers of *Cercis* and *Bauhinia* sequences. For example, there are 425 clades with ≥2 *Cercis* sequences and 1 *Bauhinia* sequence. The last column indicates the proportion of clades with a duplication pattern consistent with WGD having occurred prior to the *Cercis*-*Bauhinia* speciation, e.g. (B, (B, C) or (C, (B, C)), as opposed to a speciation pattern, e.g. ((B, B), C) or (B, (C, C)).

# of *Cercis* seqs. in clade	# of *Bauhinia* seqs. in clade	total # of clades detected	# of clades labelled as duplication at MRCA	percent of duplication clades
≥2	≥2	249	212	85%
≥2	1	425	183	43%
1	≥2	3205	2036	63%

A feature readily apparent in the phylogeny is its division into three clades: one with sequences marked “CYC1” (salmon), one with sequences marked “CYC2” (orange), and one unlabeled (red) (Fig 4.5). Most species have two representatives in the phylogeny: one in the CYC1 clade and one in the CYC2 clade – except in *Cercis* (three species), for which only one sequence was amplified (or recovered from the genome assembly, in the case of *C. canadensis*). Although the favored topology places *Cercis* sequences sister to sequences from other Cercidoideae, bootstrap support for this relationship is weak. Alternative resolutions thus are not ruled out, including placement of the *Cercis* clade sister to either CYC1 or CYC2. This would be consistent with the pattern observed in the trees in Fig 4.3, i.e., [(C, B1), B2] – and would be consistent with a model of allopolyploidy (see section “Discussion”).

Chromosome Counts Across the Legume Phylogeny

Phylogenetic and chromosome count data can be combined in order to explore chromosomal evolution across the legumes. We combined the extensive matK-based phylogeny from the LPWG [1], with count data from the Chromosome Counts Database (CCDB version 1.45) [48]. The CCDB contains 27,947 count reports for legume species, spanning 477 genera. For many genera, there are numerous reports; for example, Acacia has 472 reported counts across 152 species. We determined the modal gametic chromosomal count value, "n," for each genus (for example, in Acacia, the modal count is n = 13, of the 152 species with counts, 71% have n = 13). We then displayed these modal counts on the species phylogeny, using one species as the representative for each genus in the phylogeny.

In Fig 4.6, 4.7, a partially collapsed phylogeny has been annotated and summarized for ease of presentation. Some particularly well-represented clades have been collapsed; for example, the mimosid clade contains 47 species with chromosomal counts; these have been collapsed in Fig 4.7, and the overall modal count for that clade is presented as an annotation (the mode for the chromosomal count is n = 14 for the mimosoid clade within the Caesalpinioideae). See Table 4.5 for counts in each clade.

At the subfamily level, the modal chromosome counts are generally unambiguous, with the exception of the Papilionoideae, with a more complex pattern of chromosome counts. The Papilionoideae, being an unusually large subfamily (containing ~13,800 species in that subfamily and more than 70% of legume species; [49]), has been treated in a separate analysis [50]. However, we note here that the groups sister to the large crown clades of papilionoid species, e.g., *Swartzia*, *Myroxylon*, and *Cladrastis*, have 13 and 14 as the most frequent counts (Fig 4.6 and Table 4.5). The clades of the crown group generally have lower counts: 11 for *Amphimas*, *Holocalyx*, *Andira* dispersed along the grade with the genistoid, dalbergioid, and

baphioid clades. Among the remaining papilionoid clades (containing the majority of species in the subfamily), chromosome counts are varied, but are generally in the range of 7–11 chromosomes.

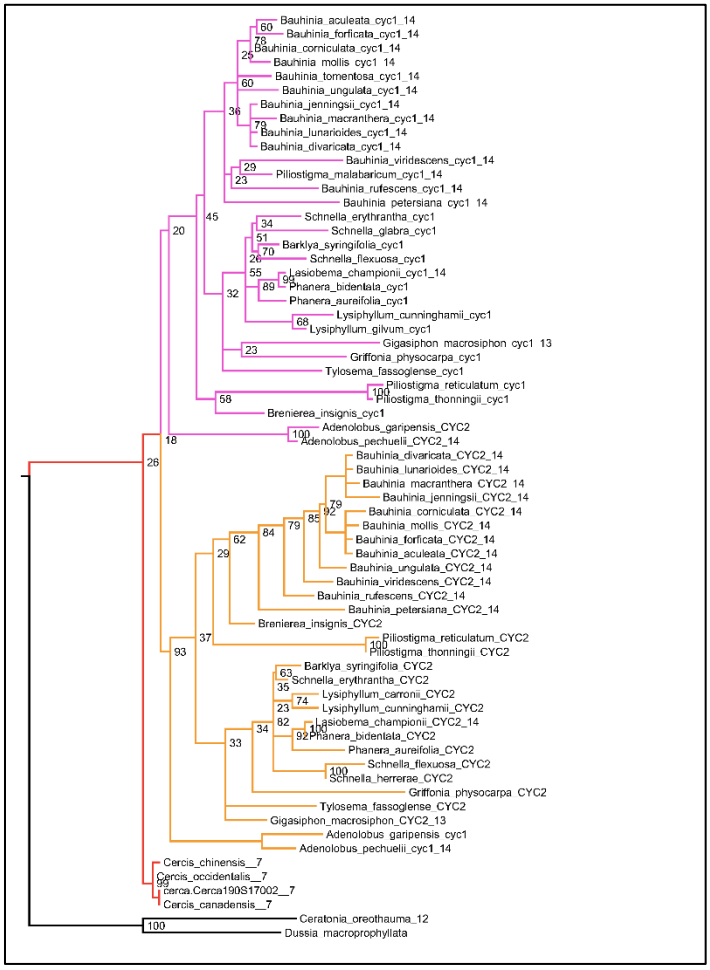

Fig 4.5. CYCLOIDEA gene tree, for species in subfamily Cercidoideae. For all species but *Cercis* (red), there are two gene copies: in the clades labeled “CYC1” (pink) and “CYC2” (orange). Where chromosome counts are available, the haploid count is indicated at the end of the sequence identifier. These values are 7 for the three included *Cercis* species, and 14 for all other species for which counts have been determined within the Cercidoideae, save *Gigasiphon macrosiphon*, which has 13. For *C. canadensis*, one sequence has been amplified using PCR and one sequence (Cerca190S17002) comes from the genomic assembly. One of several possible

rootings is shown (with bootstrap support values indicated), based on comparison with CYCLOIDEA orthologs from *Ceratonia oreothauma* (carob relative, from the Caesalpinioideae) and *Dussia macroprophyllata* (an early-diverging species from the Papilionoideae).

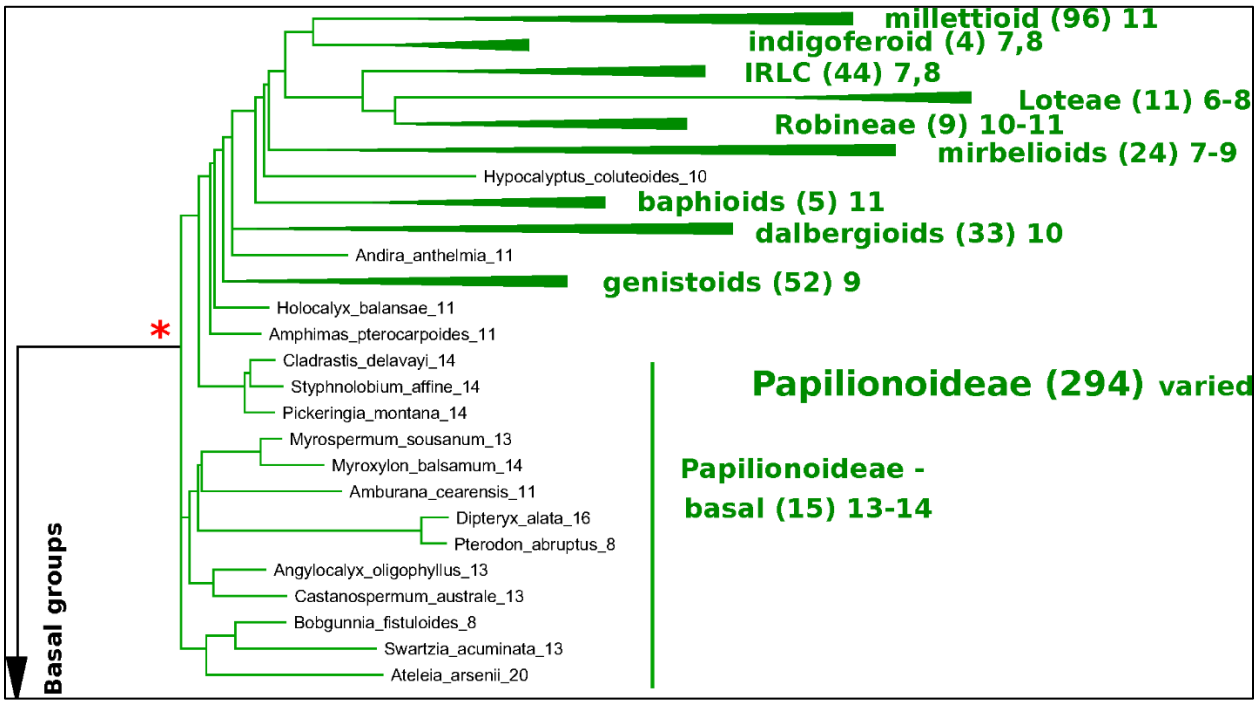

Fig 4.6. Papilionoid portion of the matK-based species phylogeny for representative species in the legumes, with chromosome count data (Fig 4.6, 4.7). matK-based species phylogeny for representative species in the legumes (derived from Legume Phylogeny Working Group et al., 2017), with chromosome count data. Only species for which chromosome counts are available are shown, with the exception of the Cercidoideae (Fig 4.7), where additional species are shown for context in that subfamily. Chromosomal counts are given as the mode for the indicated genus, where there are differences in the genus. Some particularly well-represented clades have been collapsed and are represented by a colored triangle. The number of genera with counts is given in parentheses – for example, 96 genera are represented in the triangle representing the millettioid clade (top of Fig 4.6), and 47 genera are represented in the triangle representing the

Mimosoid clade (top of Fig 4.7). Red asterisks indicate polyploidy events – either known (e.g., Papilionoideae) or hypothesized (e.g., Dialioideae).

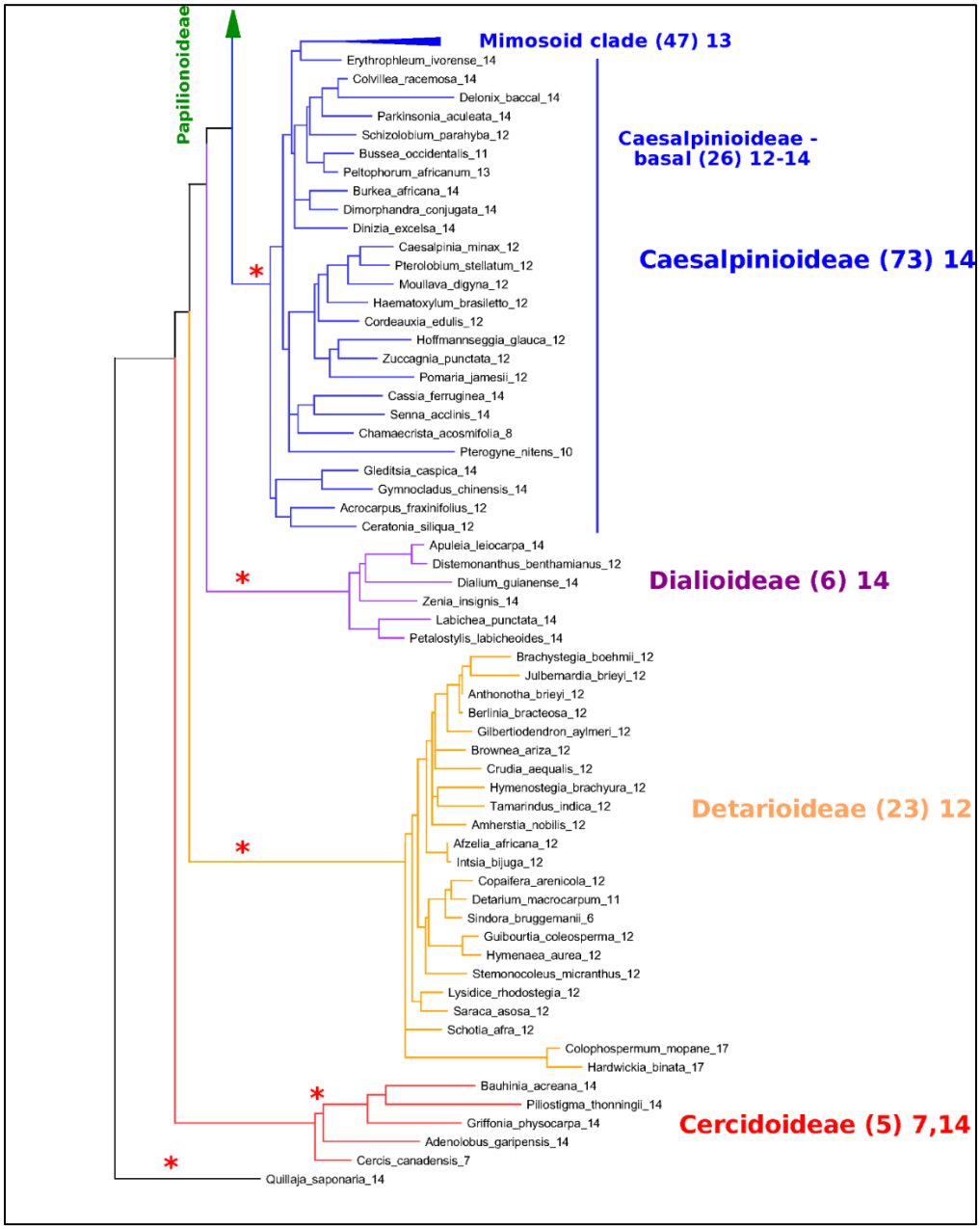

Fig 4.7. Non-papilionoid portion of the matK-based species phylogeny for representative species in the legumes, with chromosome count data. Fig 4.7 extends Fig 4.6; see description under Fig 4.6. The relative placements of the subfamilies are uncertain, with the Cercidoideae and Detarioideae, best considered as a polytomy, given current phylogenetic

resolutions (Legume Phylogeny Working Group et al., 2017). *Indicate polyploidy event – either known (e.g., Papilionoideae) or hypothesized (e.g., Dialioideae).

Table 4.5. Counts of genera with indicated 1n haploid (gametic) chromosome numbers, by subfamily or clade. Each cell (except for the count summaries in the last three columns) contains the number of genera with a 1n chromosome count indicated (column), for that clade (row). For example, in the Caesalpinioideae (which includes the mimosoid clade), 31 genera have a chromosome count of 13. (For most genera, all species have the same chromosome count, but where count differences are reported in the literature, the modal value is used for the genus). For each clade, the most frequent chromosome count is highlighted in bold, and the most frequent count values are listed on the right.

Clade \ Count	6	7	8	9	10	11	12	13	14	15	16	>16	total	frequent
Papilionoid - derived	4	21	**57**	**36**	**39**	**77**	6	0	5	0	6	27	278	8-11
Papilionoid - grade	0	0	0	0	0	**3**	0	0	0	0	0	0	3	11
Papilionoid - early	0	0	2	0	0	1	0	**4**	**4**	0	1	1	13	13-14
Caesalp - mimosoid	0	0	1	0	0	0	1	**31**	5	0	0	3	41	13
Caesalp - early	0	0	1	0	1	1	**13**	**4**	**12**	0	0	0	32	12-14
Dialidoideae	0	0	0	0	0	0	1	0	**5**	0	0	0	6	14
Detarioideae	1	0	0	0	0	1	**19**	0	0	0	0	2	23	12
Cercidoideae	0	**1**	0	0	0	0	0	0	**4**	0	0	0	5	7, 14

The Caesalpinioideae has generally clear count patterns: 14 for the large mimosoid clade and 12–14 for the remaining, early-diverging taxa (Table 4.5). Across 73 genera with counts in the Caesalpinioideae, 66 have modes at n = 12, 13, or 14 (14, 35, 17, respectively – combining "early" and "mimosoid" in Table 4.5). There are some intriguing exceptions, however; for example, *Calliandra* and *Chamaecrista* and have n = 7–8, despite being nested in clades with n =

13 or 14 – apparently indicating chromosomal fusions or reductions of some sort; and other genera such as *Neptunia* and *Leucaena*, have n = 28 and 52, respectively, suggesting ploidy increases from n = 14 and 13.

For the Dialioideae, five of six genera with count data have n = 14. For the Detarioideae, 19 of 23 genera with count data have n = 12. For the Cercidoideae, four genera (*Bauhinia*, *Piliostigma*, *Griffonia*, and *Adenolobus*) with count data have n = 14, and only *Cercis* has n = 7. The nearest outgroup species to the legumes may also be informative. *Quillaja saponaria* (Quillajaceae) which shows evidence of a WGD (via transcriptome K_s data; Cannon et al., 2015 [4]), has n = 14. Another near outgroup, *Suriana maritima* (Surianaceae), has n = 9; its WGD status is not known directly, though it lacks duplication in any of its CYC-like genes [51].

Genome Sizes in the Cercidoideae

Roberts and Werner (2016) [52] report an average of 2C = 0.751 pg for 30 accessions across 9 *Cercis* species. Using the conversion ratio of 1 pg = 978 Mb [53], this gives a *Cercis* genome size estimate of 1C = 0.751 pg * (978 Mb / 1 pg) / 2 = 367 Mbp. This compares with reported 1C genome sizes for several *Bauhinia* species: 573 Mbp for *B. purpurea*; 613 Mbp for *B. tomentosa*, and 620 Mbp for *Lysiphyllum hookeri* (formerly *B. hookeri*) [54]. These values are ~1.5 to ~1.6 times larger than *Cercis* – which is consistent with the *Bauhinia* genomes having doubled relative to *Cercis* (followed by moderate increase in *Cercis* and/or decrease in *Bauhinia* – or a situation of an allopolyploid *Bauhinia* being derived from two genomes of different sizes – one contributed by a *Cercis* progenitor and one presumably now extinct). A size of 381 Mbp for *Cercis* is also small relative to other reported legume genomes; for example, the estimated sizes of *L. japonicus*, *M. truncatula*, *P. vulgaris*, and *C. arietinum*, respectively, are 472–597 Mbp,

465–562 Mbp, 587–637, 738–929 [55–59]. Indeed, in comparison with genome size reports for 722 legume species and 84 genera from the Kew C-value database [60]), the *Cercis* estimate of n = 367 Mbp would be smaller than all but one other legume genome (Lablab niger also has an estimated size of 367 Mbp). For all reported legume genera (taking median value per genus where values are available for multiple species in a genus), the average haploid genome size is 1,424 Mbp and the median is 1,157 Mbp

Discussion

This study examines evidence regarding ploidy in the legume family, particularly focusing on subfamily Cercidoideae. What motivates this focus is the hypothesis that *Cercis*, sister to the remainder of the Cercidoideae, has no history of polyploidy – which may be in contrast to all other legume species. This would make *Cercis* valuable as a genomic model for the legumes, and would also help to clarify histories of chromosome evolution throughout the rest of the large and diverse legume family. Specifically, if *Cercis* did not undergo a WGD relative to the common ancestor of legumes, and if the ancestors of other lineages in the Cercidoideae, Dialioideae, Detarioideae, Caesalpinioideae, and Papilionoideae did, then the legume clade as a whole is not fundamentally polyploid relative to its sister taxa. Combined with evidence that the papilionoid WGD affects all papilionoid species but does not extend to species in the caesalpinioid or detarioid subfamilies [4], the necessary inference is that there must have been multiple, independent events: at a minimum, one in the Cercidoideae and another in the Papilionoideae – and our findings here are also consistent with our previous conclusion of independent polyploidy events early in the Caesalpinioideae and Detarioideae [4]. We have no information about ploidy in the monogeneric Duparquetioideae; and it is not known directly whether species in the Dialioideae experienced a WGD, though chromosome counts of 12–14 in

Dialioideae are consistent with the hypothesis that they too are polyploid. The cumulative evidence that *Cercis* lacks a legume-era WGD is substantial. Recapping:

- In K_s plots (Fig 4.1, 4.2), there is no peak indicating WGD in *Cercis* – particularly, in plots derived from synteny comparisons. In contrast, such peaks are clearly evident in diverse legume lineages including *Phaseolus*, *Bauhinia*, and *Chamaecrista*. While there is no such peak in the *Cercis* self-comparison, there are clear peaks in comparisons of *Cercis* to each of the other species examined, indicating that the lack of K_s peak is not due to something essentially wrong with gene-calls in *Cercis* (the gene calls have homologs with the comparison legume species, and those homologs can be aligned in-frame with those homologs, giving reasonable K_s results).
- In genomic synteny comparisons between *Cercis*, *Phaseolus*, and *Prunus* (the latter two with known duplication histories), the duplication status of *Cercis* looks like that of *Prunus* rather than *Phaseolus* – i.e., lacking a WGD in the timeframe of the fabidae.
- In phylogenomic analyses of 14,709 gene-family trees (Table 4.3), sequence counts aggregated across all trees show a pattern consistent with at least one WGD in each species examined except *Cercis*. Examining the proportion of gene families with two or more sequences for a species to families with only one sequence, all species examined have a ratio ranging from 54 to 80% (and 632% for *G. max*, which had an additional recent WGD), in contrast to 24% for *Cercis*. For comparison, this ratio is 69% in the set of 177 conserved collinear genes in the triplicated *B. oleracea* genome segments identified by Town et al. (2006) [61].

- Mining the gene families for phylogenetic topologies within the Cercidoideae (Table 4.4), the overwhelming majority of clades have a pattern of two *Bauhinia* sequences to one *Cercis* sequence (roughly tenfold more frequently than the other options combined).
- Diverse species within the Cercidoideae all show a pattern of duplicated *CYCLOIDEA*-family genes, with the exception of *Cercis*, which has only one *CYCLOIDEA* gene – whether assayed through amplification with degenerate primers for *CYCLOIDEA*, or through gene prediction in the *Cercis* genomic sequence (Fig 4.5). All phylogenetic analyses (whether based on plastid or nuclear sequences) resolve *Cercis* as sister to the remainder to Cercidoideae, in line with a WGD after the split with *Cercis* (although rooting in Fig 4.5 is uncertain, so *Cercis* could group with one or the other of the *CYCLOIDEA* gene forms in the gene family).
- A survey of chromosome count data for 477 legume genera, examined in a phylogenetic context (Fig 4.7, Table 4.5), shows a pattern consistent with WGDs affecting all subfamilies and most genera – with the exception of *Cercis* itself. Models in which most legumes are polyploid have been proposed in earlier studies [62, 63], on the basis of chromosome numbers. In the Cercidoideae, the most frequent chromosome count is n = 14 for most species, but 7 in *Cercis*; in the Detarioideae, the modal chromosome count is 12; in the Dialioideae, the modal count is 14; in the Caesalpinioideae, the modal count is 14; and in the Papilionoideae, the modal count for early-diverging genera (e.g., *Swartzia, Angylocalyx, Cladrastis*), the most common counts are 13 and 14. Crown-group clades have highly variable counts (generally in the range of 7–11 chromosomes), so we hypothesize a doubling from 7 to 14 leading to the papilionoid origin, then a reduction

from 14 to lower numbers for crown-group clades (dalbergioids, baphioids, mirbelioids, Robineae, Loteae, IRLC, indigoferoid, and millettioid).

- Genome sizes in the Cercidoideae are consistent with WGD in *Bauhinia* and not *Cercis*. The *Cercis* genome is approximately 367 Mbp, while values for *Bauhinia* species range from 573 to 620 Mbp. A *Cercis* genome size of 367 Mbp is tied for smallest in the legume family, and is less than a third the median reported genome size of 1,157 Mbp, across 84 legume genera. We note this result with a caveat, however, that genome sizes can be highly variable, even within a single genus – affected by mechanisms such as bursts of transposon expansions – e.g., variations in *Nicotiana* [64] or in *Aeschynomene* [65].

Further analyses of evolutionary changes due to the differing WGD status between *Cercis* and other legumes will be of interest – both at the fine scale (e.g., determining the fate of duplicated genes in various lineages, relative to *Cercis*) and at larger structural scales (e.g., determining structural changes in chromosomes following several independent WGD events) These comparisons would benefit from improved assemblies and annotations, spanning a broader range of legume clades. For example, we expect both *Chamaecrista* (as a nodulator in the Mimosoideae) and *Cercis* (as an early-diverging non-nodulator) to be useful in better understanding the origin and evolution of nodulation symbioses – as investigated in several recent papers [18, 66, 67].

An initially puzzling result from our analysis was the fact that the K_s peak for the *Bauhinia* self-comparison (*Bauhinia*–*Bauhinia*) appears significantly "older" than the *Bauhinia*–*Cercis* speciation peak, at 0.25 and 0.15, respectively (Fig 4.1A). Similarly, most gene tree

topologies (63%) that have two or more *Bauhinia* sequences and one *Cercis* sequence (Table 4.4, row 3) have a configuration of (B, (B, C)), indicating duplication prior to speciation – in contrast to what might be expected given a simple model of *Cercis–Bauhinia* speciation followed by WGD in *Bauhinia*. In the latter case, the expected pattern would be [(B, B), C] – which is observed in the minority of cases (37%). We note that an apparent speciation pattern may be due either to a WGD or to local, private duplications. Private duplications are common in plant genomes. For example, in *M. truncatula*, more than a third of paralogs are derived from local duplications [43]. However, local duplications tend to be evident in K_s plots as a recent peak, with maximum near zero – as is seen, for example, in the *Phaseolus–Phaseolus* comparison in Fig 4.1. This is the typical pattern described by Lynch and Conery (2000) [68] for eukaryotes generally. The results of our phylogenetic pattern-mining tests are consistent with what we observe (albeit anecdotally) in visual inspection of many trees, exemplified by Fig 4.3, in which there is a duplication of the *Bauhinia* paralogs in both trees, apparently followed by orthologous split between one of the *Bauhinia* sequences and the *Cercis* sequence.

A model that could accommodate the K_s and tree-topology results is one of allopolyploidy, in which a progenitor of *Cercis* speciated to give another (perhaps now-extinct) diploid species (Fig 4.8A). These species diverged for some time, and then the two species contributed their genomes to a new allopolyploid species that was the progenitor of the remaining Cercidoideae. Following allopolyploidy, the two lineages (diploid *Cercis* and polyploid *Bauhinia*) would then have proceeded to diverge and diversify – *Cercis* more slowly and the remaining species in Cercidoideae more rapidly. The current gene family view would then be as observed in e.g., Fig 4.3, or in the model in Fig 4.8B.

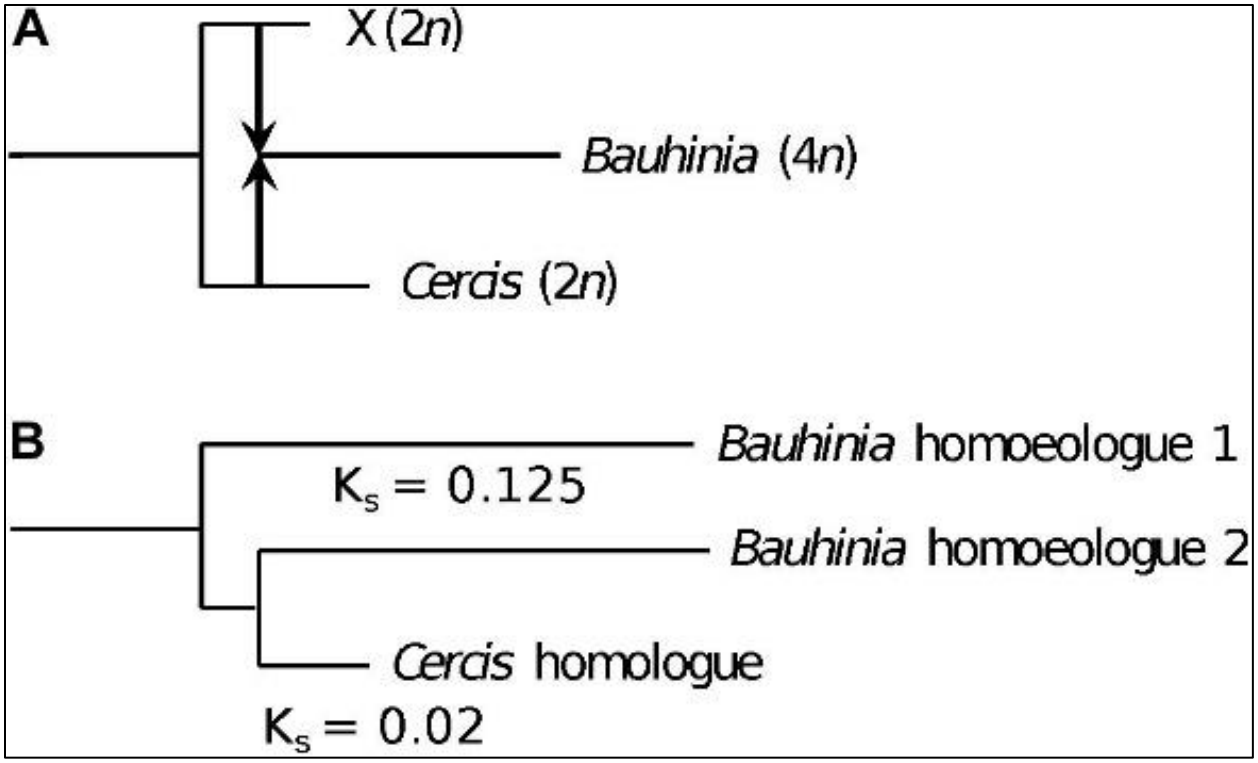

Fig 4.8. Allopolyploid origin of *Bauhinia*. (A) Species history, showing divergence between two diploid (2*n*) species: (1) the ancestor of *Cercis* and (2) a second species that became extinct ("X"). At some point after the species divergence, the two diploid species hybridized (arrows), followed by genome doubling to produce the allopolyploid (4*n*) ancestor of *Bauhinia* (and other Cercidoideae). **(B)** Representative gene tree sampled from *Bauhinia* and *Cercis*, showing the relationships of the single homologous gene in *Cercis* to the two homoeologs in allopolyploid *Bauhinia*. The *Bauhinia* homoeolog 2, contributed by the *Cercis* ancestor, is sister to the *Cercis* gene. The Cercis gene has a K_s of ~0.145 compared with the *Bauhinia* homeolog 2; and each *Bauhinia* homoeolog has a K_s of 0.25 with respect to the other *Bauhinia* homoeolog. The relationship between the species history and the gene tree is complicated by the hypothesized slower substitution rate in Cercis.

A model that could accommodate the K_s and tree-topology results is one of allopolyploidy, in which a progenitor of *Cercis* speciated to give another (perhaps now-extinct) diploid species (Fig 4.8A). These species diverged for some time, and then the two species contributed their genomes to a new allopolyploid species that was the progenitor of the remaining Cercidoideae. Following allopolyploidy, the two lineages (diploid *Cercis* and polyploid *Bauhinia*) would then have proceeded to diverge and diversify – *Cercis* more slowly and the remaining species in Cercidoideae more rapidly. The current gene family view would then be as observed in e.g., Fig 4.3, or in the model in Fig 4.8B.

Precedent for a significant period of species divergence prior to allopolyploidy is seen, for example, in *Arachis*: the allopolyploid *A. hypogaea* was formed, within about the last 10 thousand years, from the merger of *A. duranensis* and *A. ipaensis*, which diverged an estimated 2.16 Mya [10]. Another similar example is in cotton, where the allotetraploid *Gossypium hirsutum* L. is a merger of genomes from progenitor species similar to the extant diploid species *G. ramondii* Ulbrich and *G. herbaceum* L. [69–71] In this case, the diploid species diverged c. 5–10 Mya and merged to form *G. hirsutum* c. 1–2 Mya [69, 72].

The genus *Cercis* contains 10 species and all phylogenetic analyses to date have supported the genus as monophyletic. This is a well-defined group of north temperate trees (North America, Eurasia and eastern Asia). All species for which counts are available are diploid[2]. There appears to be relatively low genetic diversity within the genus based on plastid and nuclear ribosomal ITS sequences [73, 74]. *C. chingii* (n = 14) is resolved as sister to the other species in the genus in the studies by Davis et al. (2002) [73], and differs from the other species by its coriaceous, unwinged, dehiscent fruit. The other species are morphologically quite

similar. It's not clear if one of the present day *Cercis* species could better represent an ancestral parental genome resulting in the whole genome duplication.

Cercis genes do appear to have evolved remarkably slowly (at least in the sense of accumulating point mutations that affect K_s and branch lengths). A tree calculated by algebraically solving evolutionary "distance paths" along a gene tree (Fig 4.1, 4.2, lower right), using K_s-based branch lengths, shows a *Cercis* evolutionary rate less than a quarter that of *Bauhinia*, and roughly a tenth that of *Phaseolus* since the papilionoid WGD. The slow *Cercis* rate is also evident in many gene family trees, such as the two shown in Fig 4.3. The *matK* gene tree also shows remarkably short branches for *Cercis*. It is conceivable that the slower evolutionary rate seen in *Cercis* than other legumes might be partly due to the lack of WGD-derived "extra" genes in *Cercis* –perhaps presenting extra evolutionary constraints than for duplicated genes. The outcrossing, long-lived tree form might also constrain evolutionary rates (injecting older gametes into new progeny) – although of course these conditions are shared with many species.

Conclusion

The evidence from diverse sources indicates that *Cercis* may be unique among legume lineages in lacking any evidence for a WGD; that its last duplication event was probably the eudicot "gamma" triplication event; that the genomes of other Cercidoideae and all other legume subfamilies are likely to have been shaped by independent WGD events; that the most likely model for WGD and speciation timing in the Cercidoideae is allopolyploidy – with a *Cercis* progenitor contributing one subgenome to the allopolyploid *Bauhinia* progenitor; and lastly, that *Cercis* has evolved at a strikingly low rate since its divergence from other Cercidoideae. Taken

together, these findings suggest that *Cercis* may serve as a useful genomic model for the legumes, likely representing the duplication status of the progenitor of all legumes.

References

1. Azani N, Babineau M, Bailey CD, et al (2017) A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). TAXON 66:44–77

2. Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. Syst Biol 54:575–594

3. Bruneau A, Mercure M, Lewis GP, Herendeen PS (2008) Phylogenetic patterns and diversification in the caesalpinioid legumesThis paper is one of a selection of papers published in the Special Issue on Systematics Research. Botany 86:697–718

4. Cannon SB, McKain MR, Harkess A, et al (2015) Multiple Polyploidy Events in the Early Radiation of Nodulating and Nonnodulating Legumes. Mol Biol Evol 32:193–210

5. Lewis GP (2005) Legumes of the World. Royal Botanic Gardens Kew

6. Sinou C, Forest F, Lewis GP, Bruneau A (2009) The genus *Bauhinia* s.l. (Leguminosae): a phylogeny based on the plastid trnL–trnF region. Botany 87:947–960

7. Wang Y-H, Wicke S, Wang H, Jin J-J, Chen S-Y, Zhang S-D, Li D-Z, Yi T-S (2018) Plastid Genome Evolution in the Early-Diverging Legume Subfamily Cercidoideae (Fabaceae). Front Plant Sci 9:138

8. Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 181:1–20

9. Jiao Y, Leebens-Mack J, Ayyampalayam S, et al (2012) A genome triplication associated with early diversification of the core eudicots. Genome Biol 13:R3

10. Bertioli DJ, Cannon SB, Froenicke L, et al (2016) The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat Genet 48:438–446

11. Varshney RK, Chen W, Li Y, et al (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

12. Schmutz J, Cannon SB, Schlueter J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

13. Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

14. Kang YJ, Kim SK, Kim MY, et al (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. Nat Commun 5:1–9

15. Sato S, Nakamura Y, Kaneko T, et al (2008) Genome Structure of the Legume, *Lotus japonicus*. DNA Res 15:227–239

16. Tang H, Krishnakumar V, Bidwell S, et al (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics 15:312

17. Varshney RK, Song C, Saxena RK, et al (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240–246

18. Griesmann M, Chang Y, Liu X, et al (2018) Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science. https://doi.org/10.1126/science.aat1743

19. Cannon SB, McKain MR, Harkess A, et al (2015) Multiple Polyploidy Events in the Early Radiation of Nodulating and Nonnodulating Legumes. Mol Biol Evol 32:193–210

20. Verde I, Abbott AG, Scalabrin S, et al (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 45:487–494

21. Phytozome v12.1: Info. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Csativus. Accessed 26 Nov 2019

22. Jaillon O, Aury J-M, Noel B, et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. nature 449:463

23. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E (2015) The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. genesis 53:474–485

24. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641

25. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421

26. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586–1591

27. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

28. Tang H (2019) tanghaibao/bio-pipeline.

29. Cock PJA, Antao T, Chang JT, et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423

30. Larkin MA, Blackshields G, Brown NP, et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

31. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–W612

32. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584

33. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

34. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 41:e121–e121

35. Stamatakis A, Hoover P, Rougemont J (2008) A Rapid Bootstrap Algorithm for the RAxML Web Servers. Syst Biol 57:758–771

36. Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E (2017) SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics 33:2197–2198

37. Felsenstein J (1993) PHYLIP (phylogeny inference package), version 3.5 c. Joseph Felsenstein.

38. Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11:24

39. Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33:1635–1638

40. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T (2007) The human phylome. Genome Biol 8:R109

41. Cui L, Wall PK, Leebens-Mack JH, et al (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738–749

42. Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

43. Young ND, Debellé F, Oldroyd GED, et al (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524

44. Tang H, Lyons E, Pedersen B, Schnable JC, Paterson AH, Freeling M (2011) Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics 12:102

45. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J (2010) Genome sequence of the palaeopolyploid soybean. nature 463:178

46. Citerne HL, Luo D, Pennington RT, Coen E, Cronk QCB (2003) A Phylogenomic Investigation of CYCLOIDEA-Like TCP Genes in the Leguminosae. Plant Physiol 131:1042–1053

47. Citerne HL, Pennington RT, Cronk QCB (2006) An apparent reversal in floral symmetry in the legume Cadia is a homeotic transformation. Proc Natl Acad Sci 103:12017–12020

48. Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM, Salman-Minkov A, Mayzel J, Chay O, Mayrose I (2015) The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. New Phytol 206:19–26

49. Cardoso D, Queiroz LP de, Pennington RT, Lima HC de, Fonty É, Wojciechowski MF, Lavin M (2012) Revisiting the phylogeny of papilionoid legumes: New insights from comprehensively sampled early-branching lineages. Am J Bot 99:1991–2013

50. Ren L, Huang W, Cannon SB (2019) Reconstruction of ancestral genome reveals chromosome evolution history for selected legume species. New Phytol 223:2090–2103

51. Zhao Z, Hu J, Chen S, Luo Z, Luo D, Wen J, Tu T, Zhang D (2019) Evolution of CYCLOIDEA-like genes in Fabales: Insights into duplication patterns and the control of floral symmetry. Mol Phylogenet Evol 132:81–89

52. Roberts DJ, Werner DJ (2016) Genome Size and Ploidy Levels of *Cercis* (Redbud) Species, Cultivars, and Botanical Varieties. HortScience 51:330–333

53. Dolezel J, Bartos J, Voglmayr H, Greilhuber J (2003) Nuclear DNA content and genome size of trout and human. Cytom Part J Int Soc Anal Cytol 51:127–8; author reply 129

54. Bennett MD, Leitch IJ (2005) Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects. Ann Bot 95:45–90

55. Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Report 9:208–218

56. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227–239

57. Bennett MD, Leitch IJ (2011) Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. Ann Bot 107:467–590

58. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240

59. Tang H, Krishnakumar V, Bidwell S, et al (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics 15:312

60. Bennett M, Leitch I (2012) Plant DNA C-values database (release 6.0, Dec. 2012). WWW Doc. URL Httpdata Kew Orgcvaluesaccessed 14 Oct. 2014

61. Town CD, Cheung F, Maiti R, et al (2006) Comparative Genomics of *Brassica oleracea* and *Arabidopsis thaliana* Reveal Gene Loss, Fragmentation, and Dispersal after Polyploidy. Plant Cell 18:1348–1359

62. Goldblatt P (1981) Chromosome Numbers in Legumes II. Ann Mo Bot Gard 68:551–557

63. Doyle JJ (2012) Polyploidy in Legumes. In: Soltis PS, Soltis DE (eds) Polyploidy Genome Evol. Springer, Berlin, Heidelberg, pp 147–180

64. Leitch IJ, Hanson L, Lim KY, Kovarik A, Chase MW, Clarkson JJ, Leitch AR (2008) The Ups and Downs of Genome Size Evolution in Polyploid Species of *Nicotiana* (Solanaceae). Ann Bot 101:805–814

65. Brottier L, Chaintreuil C, Simion P, et al (2018) A phylogenetic framework of the legume genus Aeschynomene for comparative genetic analysis of the Nod-dependent and Nod-independent symbioses. BMC Plant Biol 18:333

66. Battenberg K, Potter D, Tabuloc CA, Chiu JC, Berry AM (2018) Comparative Transcriptomic Analysis of Two Actinorhizal Plants and the Legume *Medicago truncatula* Supports the Homology of Root Nodule Symbioses and Is Congruent With a Two-Step Process of Evolution in the Nitrogen-Fixing Clade of Angiosperms. Front Plant Sci. https://doi.org/10.3389/fpls.2018.01256

67. Velzen R van, Holmer R, Bu F, et al (2018) Comparative genomics of the nonlegume Parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses. Proc Natl Acad Sci 115:E4700–E4709

68. Lynch M, Conery JS (2000) The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151–1155

69. Wendel JF (1989) New World tetraploid cottons contain Old World cytoplasm. Proc Natl Acad Sci 86:4132–4136

70. Flagel LE, Wendel JF, Udall JA (2012) Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genomics 13:302

71. Paterson AH, Wendel JF, Gundlach H, et al (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492:423–427

72. Fang L, Wang Q, Hu Y, et al (2017) Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet 49:1089–1098

73. Davis CC, Fritsch PW, Li J, Donoghue MJ (2002) Phylogeny and Biogeography of *Cercis* (Fabaceae): Evidence from Nuclear Ribosomal ITS and Chloroplast ndhF Sequence Data. Syst Bot 27:289–302

74. Coşkun F, Parks CR (2009) A molecular phylogenetic study of red buds (*Cercis* L., Fabaceae) based on ITS nrDNA sequences. Pak J Bot 41:1577–1586

CHAPTER 5. FAMILY-SPECIFIC GAINS AND LOSSES OF PROTEIN DOMAINS IN LEGUME AND GRASS PLANT FAMILIES

Akshay Yadav, David Fernández-Baca, Steven B. Cannon

Modified from a manuscript to be submitted to a peer reviewed journal

Abstract

Protein domains can be regarded as sections of protein sequences capable of folding independently and performing specific functions. In addition to amino-acid level changes, protein sequences can also evolve through domain shuffling events like domain insertion, deletion, or duplication. The evolution of protein domains can be studied by tracking domain changes in a selected set of species with known phylogenetic relationships. Here, we conduct such an analysis by defining domains as "features" or "descriptors," and considering the species (target + outgroup) as instances or data-points in a data matrix. We then look for features (domains) that are significantly different between the target species and the outgroup species. We study the domain changes in two large, distinct groups of plant species: legumes (Fabaceae) and grasses (Poaceae), with respect to selected outgroup species. We evaluate four types of domain feature matrices: domain content, domain duplication, domain abundance, and domain versatility. The four types of domain feature matrices attempt to capture different aspects of domain changes through which the protein sequences may evolve - i.e. via gain or loss of domains, increase or decrease in the copy number of domains along the sequences, expansion or contraction of domains, or through changes in the number of adjacent domain partners. All the feature matrices were analyzed using feature selection techniques and statistical tests in order to select protein domains that have significantly different feature values in legumes and grasses. We

report the biological functions of the top selected domains from analysis all the feature matrices. In addition, we also perform domain-centric Gene Ontology (dcGO) enrichment analysis on all selected domains from all 4 feature matrices to study the Gene Ontology terms associated with the significantly evolving domains in legumes and grasses. Domain content analysis revealed a striking loss of protein domains from the Fanconi Anemia (FA) pathway, the pathway responsible for the repair of interstrand DNA crosslinks. The abundance analysis of domains found in legumes revealed an increase in glutathione synthase enzyme, an antioxidant required from nitrogen fixation, and a decrease in xanthine oxidizing enzymes, a phenomenon confirmed by previous studies. In grasses, the abundance analysis showed increases in domains related to gene silencing which could be due to polyploidy or due to enhanced response to viral infection. We provide a docker container that can be used to perform this analysis workflow on any user-defined sets of species, available at https://cloud.docker.com/u/akshayayadav/repository/docker/akshayayadav/protein-domain-evolution-project.

Introduction

Protein domains are independent evolutionary units of proteins that enable proteins to evolve in a modular fashion through domain insertion, deletion, duplication, or substitution, in addition to evolution through point mutations [1, 2]. In this ability of protein domains to fold and function independently of other domains, they can be considered as “lego bricks” that can be recombined in various ways to build new proteins [3, 4]. Small proteins are usually made up of just one domain whereas large proteins are formed by combinations of multiple domains [5]. Roughly two-thirds of the prokaryotic proteins and four-fifths of the eukaryotic proteins are multi-domain proteins that are formed through recombination of two or more domains [6, 7]. The “combinability” of domains makes them prime candidates for studying evolution - both of

proteins and of species. For example, protein domains have been used to study evolution on genome-wide and species-wide scales by examining the protein domain content of the species [8–10]. Protein domain content is defined by the presence or absence of protein domains in complete genomes of the species. The importance of protein domains in studying evolution can be verified from the ability of protein domain content in reconstructing the phylogeny of life, in comparison to trees obtained from standard phylogenetic and phylogenomic approaches that utilize information from molecular markers, gene content and gene order [10].

In this study, we examine the domain combinations present in two groups of plant species - the legumes (Fabaceae) and grasses (Poaceae), treating the protein domains as species "features" that may be present or absent in the focal species. Accordingly, a data matrix was defined with rows representing species, columns representing the protein domains and the cells containing domain feature values for the respective species. We used standard feature selection and statistical testing techniques to identify protein domains that differ between the target set of species and their respective outgroups.

Gain or loss of particular domains in a group of species can provide a means of understanding trait evolution in those species [11, 12]. Protein domains can duplicate locally, giving significantly different counts of certain domains. This may provide some useful information about functions associated with those domains [13, 14]. Counts of protein domains can also increase or decrease along with the proteins that they comprise [15]. Finally, "versatile" domains can partner with multiple different domains; and versatility values can be used to study the evolution of associated functions [3, 16, 17]. We evaluated domain evolution using these types of domain feature matrices: domain content, duplication, abundance, and versatility.

We used two types of statistical methods: Mutual-Information (MI) and non-parametric statistical tests. MI measures mutual dependence between two random variables by quantifying the amount of information communicated about one random variable from another random variable [18]. MI has been routinely used for selecting meaningful features, in classification and pattern recognition problems [19–21]. Here, we used MI to quantify the mutual dependence between domain feature values and the classification between target and outgroup species. We also employed tests for significance of differences in domain feature values between the target and outgroup species. We applied Fisher's exact tests [22] for feature matrices containing discrete values, and Wilcoxon rank-sum tests [23] for feature matrices containing continuous values.

Material and Methods

We used two sets of plant species to study the species-level changes in protein domain characteristics for a given set of target species. The first set (Table 5.1) consisted of 14 legumes (from the Papilionoideae subfamily within the legume/Fabaceae family), and 10 outgroup species defined with respect to the legumes [24–45]. The second set (Table 5.2) consisted of 10 grass species (Poaceae) and 9 outgroup species defined with respect to the grasses [27, 36, 37, 39, 40, 42–54].

All target proteomes from legumes and grasses, together with their respective outgroup proteomes, were searched against domain HMMs from the Pfam database (release 32) [55] in order to assign domains to the protein sequences. The *pfam_scan.pl* script [56] was used to assign domains to proteomes, which internally uses the *hmmscan* program from the HMMER package [57]. Subsequently, the domain assignments from target proteomes and their respective outgroup proteomes were used to calculate the four types of domain feature matrices.

Table 5.1: Legumes and legume outgroups used to study protein domain evolution in the legumes

Species	Abbrev.	Genotype	Assembly	Annot.	Publication	Source
Arachis duranensis	aradu	V14167	1	1	Bertioli et al. (2015)	PeanutBase
Arachis ipaensis	araip	K30076	1	1	Bertioli et al. (2015)	PeanutBase
Arachis hypogaea	arahy				Bertioli et al. (2015)	PeanutBase
Cajanus cajan	cajca	ICPL87119	1	1	Varshney et al. (2012)	LegumeInfo
Cicer arietinum	cicar	Frontier	1	1	Varshney et al. (2013)	LegumeInfo
Glycine max	glyma	Williams 82	2	1	Schmutz et al. (2010)	Phytozome
Lotus japonicus	lotja	MG20	3	1	Sato et al. (2008)	Phytozome
Lupinus angustifolius	lupan				Hane et al. (2017)	LegumeInfo
Medicago truncatula	medtr	A17_HM341	4	2	Tang et al. (2014)	Phytozome
Phaseolus vulgaris	phavu	G19833	2	1	Schmutz et al. (2014)	Phytozome
Trifolium pratense	tripr				De Vega (2015)	LegumeInfo
Vigna angularis	vigan	Va3.0	1	3	Kang et al. (2015)	LegumeInfo
Vigna radiata	vigra	VC1973A	6	1	Kang et al. (2014)	LegumeInfo
Vigna unguiculata	vigun	IT97K	1	1	Phytozome	Phytozome
Prunus persica	prupe	Lovell	2	2.1	IPGI (2013)	Phytozome
Vitis vinifera	vitvi	PN40024	12X	12X	Jaillon et al. (2007)	Phytozome
Cucumis sativus	cucsa		1	1	Phytozome, 2017	Phytozome
Arabidopsis thaliana	arath	Col-0	TAIR10	TAIR10	Berardini et al. (2015)	Phytozome
Solanum lycopersicum	solly	LA1589	ITAG2.4	ITAG2.4	Tomato Genome Consortium (2012)	Phytozome

Table 5.1 Continued						
Species	**Abbrev.**	**Genotype**	**Assembly**	**Annot.**	**Publication**	**Source**
Gossypium raimondii	gosra		2	2.1	Paterson et al. (2012)	Phytozome
Oryza sativa	orysa		7	7.0	Ouyang et al. (2007)	Rice Genome Annotation Project
Populus trichocarpa	poptr		3	3.1	Tuskan et al. (2006)	Phytozome
Theobroma cacao	theca		2	2.1	Motamayor et al. (2013)	Cacao Genome Project
Zea mays	zeama		6	6a	Schnable et al. (2009)	

Table 5.2: Grasses and grass outgroups used to study protein domain evolution in the grasses

Species	**Abbrev.**	**Genotype**	**Assembly**	**Annot.**	**Publication**	**Source**
Setaria italica	setit	Yugu1	2	2.2	Bennetzen JL et al. (2012)	Phytozome
Setaria viridis	setvi		2	2.1	Phytozome, 2017	Phytozome
Panicum hallii	panha	filipes	3	3.1	Phytozome, 2017	Phytozome
Panicum virgatum	panvi		5	5.1	Phytozome, 2017	Phytozome
Zea mays	zeama		6	6a	Schnable et al. (2009)	
Sorghum bicolor	sorbi		3.1	3.1.1	McCormick et al. (2017)	Phytozome
Oropetium thomaeum	oroth		1	1.0	VanBuren et al. (2015)	Phytozome
Brachypodium distachyon	bradi		3	3.1	International Brachypodium Initiative (2010)	Phytozome
Brachypodium stacei	brast		1	1.1	Phytozome, 2017	Phytozome
Oryza sativa	orysa		7	7.0	Ouyang et al. (2007)	RGAP

Table 5.2 Continued						
Species	**Abbrev.**	**Genotype**	**Assembly**	**Annot.**	**Publication**	**Source**
Arabidopsis thaliana	arath	Col-0	TAIR10	TAIR10	Berardini et al. (2015)	Phytozome
Theobroma cacao	theca		2	2.1	Motamayor et al. (2013)	Cacao Genome Project
Populus trichocarpa	poptr		3	3.1	Tuskan et al. (2006)	Phytozome
Prunus persica	prupe	Lovell	2	2.1	IPGI (2013)	Phytozome
Glycine max	glyma	Williams 82	2	1	Schmutz et al. (2010)	Phytozome
Vitis vinifera	vitvi	PN40024	12X	12X	Jaillon et al. (2007)	Phytozome
Solanum lycopersicum	solly	LA1589	ITAG2.4	ITAG2.4	Tomato Genome Consortium (2012)	Phytozome
Ananas comosus	anaco		3	3	Ming et al. (2015)	Phytozome
Musa acuminata	musac		1	1	Droc et al. (2013)	Banana Genome Hub

Calculation of Domain Feature Matrices

The domain content matrix was calculated in order to represent the presence or absence of domains in target and outgroup species. Columns of the content matrix represent individual Pfam domains and rows represent species. Each cell was assigned a value of ‘1’ if the corresponding domain was detected in the species, else the cell was a value of ‘0’. Columns with domains that were present in all the target and outgroup species were uninformative, and therefore removed.

The domain duplication matrix contains the most frequent copy number of each Pfam domain in species, which was calculated as the modal value of list all possible copy counts of that domain in the corresponding species. Then, the modal value of the list was calculated and

added to each domain column and corresponding species row. Columns with constant duplication values across all the species (target + outgroup) were removed from the matrix. Also, columns with domain duplication values ≤ 1 across all the rows were removed.

The domain abundance matrix was built to represent the abundance value of protein domains in target and outgroup species. Here, we define the abundance value of domain in each species as the proportion of protein sequences from the entire proteome, that contain the domain. The abundance value of each domain in each species is calculated using the Inverse Domain Frequency (IDF) function (eq 1) which is inspired by the Inverse Document Frequency function used in text mining and Natural Language Processing (NLP) applications.

$$IDF(S,d) = \log_2 \frac{N(S)}{N(S,d)} \qquad (1)$$

Where *N(S)* is the total number of proteins in species '*S*' and *N(S, d)* is the number of proteins containing domain '*d*' in species '*S*'

The domain versatility matrix was calculated to represent the changes in the versatility values of the domains across the species. Versatility value (eq 2) for a given domain and species combination was calculated as the reciprocal of the number of domains immediately adjacent to the given domain, in the corresponding species. Here too, the columns with constant versatility values across all species (target + outgroup) were removed from the matrix.

$$V(S,d) = \frac{1}{F(S,d)} \qquad (2)$$

Where *F(S,d)* is the number of different domains adjacent to domain '*d*' in species 'S'

Finally, all four domain feature matrices were attached with an additional "species label" column containing value '1' for target species and '0' for outgroup species.

Statistical Analysis of Domain Feature Matrices

We applied two types of statistical analyses to the domain feature matrices. The Mutual Information (MI) function (eq 3) was used to calculate the MI-score for each domain feature by comparing against the species label column. The MI quantity measures how much information, on average, is communicated in the domain feature column about the classification between target and outgroup species (species label column). Feature columns of the duplication and abundance matrices were subjected to 'L^2' normalization before application of MI scoring. The L^2 normalization technique modifies the column values such that in each column the sum of the squares will always have a maximum value of 1.

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} P(x,y)\ log \frac{P(x,y)}{P(x)P(y)} \qquad (3)$$

We also tested feature columns for significance, calculating p-values to measure the difference in domain feature values between target and outgroup species. We used Fisher's exact test to evaluate feature columns from the duplication and versatility matrices. The Fisher's exact test was applied to contingency tables built using the discrete values from each domain column and the species labels. The dimensions of the contingency tables, in case of the content matrix, were always 2×2, since each domain column can have only two possible values for each species row - whereas in case of duplication and versatility matrices, the dimensions were r×2, where 'r' is the number of discrete values observed in the corresponding domain column. The Wilcoxon rank-sum test was applied for significance testing of the domain abundance matrix due to continuous values of the domain features. The p-values obtained for domains were corrected for multiple testing using the FDR method [58]. The FDR-adjusted p-values were reported for the domains.

Results

All four types of domain feature matrices were calculated for two sets of plants - the first containing 14 legume and 10 outgroup species, and the second containing 10 grass and 9 outgroup species. For all feature matrices, we applied Mutual-Information (MI) scoring and significance testing.

Domain Content Analysis

In legumes and grasses, 13 and 55 domains, respectively, showed significant presence/absence differences relative to their respective outgroups. The results show loss of 12 domains and gain of the *SHNi-TPR* domain in legumes, and loss of 33 domains and gain of 22 domains in grasses. The Pfam domains showing the most significant gain or loss in legumes and grasses are listed in Tables 5.3 and 5.4. The gained *SHNi-TPR* domain in the legumes contains an interrupted form of the TPR repeat. The *SHNi-TPR* family includes proteins such as Sim3 (yeast), NASP(Human) and N1/N2(Xenopus), which are responsible for delivering histone proteins such as H3 to centromeric chromatin [59]. Most of the missing domains in legumes are parts of multi-domain proteins found in the Fanconi Anemia (FA) pathway. The FA pathway is responsible for maintaining the chromosomal stability through repair of interstrand DNA crosslinks in a replication-dependent manner [60]. Most of the proteins in the FA pathway form a core complex known as the FA core complex which is responsible for the ubiquitination of FANCD2 and FANCI proteins [61]. Both the proteins are then localized to the site of DNA repair along with few other proteins. The FANCI is multi-domain protein made up of 5 domains: *FANCI_S2*, *FANCI_S1*, *FANCI_HD1*, *FANCI_HD2* and *FANCI_S4.* All the 5 FANCI domains are missing in legumes, which means that the entire FANCI protein is lost in legumes. In addition, FANCD2 binding *FA_FANCE* domain [62] and the C-terminal domain of FANCL

Table 5.3: Domains gained or lost in legumes with respect to legume outgroups (top 10 by MI score).

Domain Name	MI-score	FDR-adjusted p-values	Gain-Loss (+/-) status
FANCI_S2	0.6082	0.0017	-
FANCI_S1	0.5804	0.0017	-
FANCI_HD1	0.5804	0.0017	-
SHNi-TPR	0.5781	0.0017	+
WD-3	0.5666	0.0017	-
TPMT	0.5527	0.0017	-
FANCI_HD2	0.5527	0.0017	-
FANCI_S4	0.5527	0.0017	-
FA_FANCE	0.516	0.0099	-
FANCL_C	0.4604	0.0099	-

Table 5.4: Domains gained or lost in grasses with respect to grass outgroups (top 10 by MI score).

Domain Name	MI-score	FDR-adjusted p-values	Gain-Loss (+/-) status
P_C	0.7188	0.0011	+
Mur_ligase	0.7188	0.0011	-
Glutenin_hmw	0.7188	0.0011	+
DUF1618	0.7188	0.0011	+
MFS18	0.7188	0.0011	+
SEO_C	0.7188	0.0011	-
ACCA	0.7188	0.0011	-
SEO_N	0.7188	0.0011	-
DUF1719	0.7188	0.0011	+
DUF1110	0.7188	0.0011	+

protein (*FANCL_C* domain) are also missing in legumes. The missing *WD-3* domain belongs to the family of WD-repeats region, which is approximately 100 residues long and is contained within the FANCL protein, the putative E3 ubiquitin ligase subunit of the FA core complex [63]. In addition to the domains from the FA pathway, the Thiopurine-S-methyltransferase (*TPMT*) domain was also detected as lost from the legumes. This is a cytosolic enzyme involved the catalysis of S-methylation of aromatic and heterocyclic sulfhydryl compounds, such as anticancer and immunosuppressive thiopurines [64].

Among the top 10 protein domains in grasses, 6 were detected as gained and 4 were detected as lost with respect to the grass outgroups. There were 3 domains with unknown functions - *DUF1618, DUF1719, DUF1110* and 3 domains with known functions - *P_C, Glutenin_hmw, MFS18,* that were detected as present in grasses. The *P_C* domain is present at the C terminus of plant P proteins. The P proteins in maize act as transcriptional regulators of enzymes involved in a red phlobaphene pigment-producing arm of the flavonoid biosynthesis pathway [65, 66]. The domain *Glutenin_hmw* is the high molecular subunit of glutenin protein responsible for the elastic properties of gluten. The elastomeric glutenin proteins form a network that can withstand significant deformations without breaking, and return to the original conformation when the stress is removed - the property important for making dough [67]. The Male Flower Specific protein 18 (*MFS18*) domain found in the MFS18 protein in maize is rich in glycine, proline and serine. The MFS18 mRNA is found to accumulate in vascular bundle in the glumes, anther walls, paleas and lemmas of mature florets [68].

The four domains *Mur_ligase*, *SEO_N*, *SEO_C* and *ACCA*, were among the top 10 domains detected as lost in most grasses with respect to the selected outgroups. The *Mur_ligase* domain is the catalytic domain found in the Mur ligase family of enzymes that catalyze the successive steps in the synthesis of peptidoglycan [69]. The *SEO_N* and *SEO_C* in domains are respectively found at the N and C terminus of Sieve Element Occlusion (SEO) proteins also known as phloem proteins or forisomes. These phloem proteins remain associated with cisternae of the endoplasmic reticulum of the sieve elements after differentiation and provide rapid protection against wounding of sieve tubes by forming a gel-like mass [70]. The *ACCA* domain is the alpha isoform of the carboxyltransferase subunit of Acetyl Co-A carboxylase enzyme. The *ACCA* domain is known to play an important role in production of Malonyl-CoA in fatty acid synthesis [71].

Domain Duplication Analysis

Application of MI-scoring and Fisher's exact tests on domain features of duplication matrices revealed a single domain (of unknown function) in legumes and 8 types of domains in grasses that were significantly different (FDR ≤ 0.05) in their copy numbers as compared to the copy numbers observed in their respective outgroup sets. The domain *DUF812* is present in 1 copy in all legume sequences except *Medicago*, and in 2 copies in all outgroups except rice and maize (MI-score = 0.519444; FDR = 0.000993). Among the 8 significantly different domains in grasses (Table 5.5), 4 of the domains have increased in copy numbers and 4 have decreased in copy numbers. The domains *DUF775*, *SPX*, *zf-PARP* and *FANCF* are present in 2 copies in the majority of grass sequences and in 1 copy in majority of outgroup sequences. The *SPX* domain is a 180 residue-long protein domain found at the N-terminus of a family of proteins involved in G-protein associated signal transduction [72–74]. The *zf-PARP* domain resides at the amino-

Table 5.5: Domains with significant differences in copy numbers between grasses and grass outgroups (top 10 by MI score).

Domain Name	MI-score	FDR-adjusted p-values	Gain-Loss (+/-) status
Sec39	0.6168	0.0178	-
Prenyltrans	0.5845	0.0339	-
DUF775	0.5642	0.0178	+
Nop16	0.5193	0.0356	-
SPX	0.4798	0.0356	+
zf-PARP	0.4715	0.0445	+
mTERF	0.4447	0.0356	-
FANCF	0.4329	0.0359	+

terminal region of Poly (ADP-ribose) polymerase protein, which is an important regulatory component in the cellular response to DNA damage. This domain is known to act as a DNA nick sensor [75]. The *FANCF* domain is present in the Fanconi Anemia group F protein involved in Fanconi Anemia (FA) DNA repair pathway. Inactivation of the FANCF protein induced by methylation may play an important role in occurrence of ovarian cancers [76].

The domains *Sec39*, *Prenyltrans*, *Nop16* and *mTERF* show decrease in copy numbers, with 2, 2, 3 to 5 and 2 copies in majority of the outgroup species and 1, 1, 1 to 3 and 1 copies respectively, in the majority of grasses. The *Sec39* domain is a part of "secretory pathway protein 39," which is involved in ER-Golgi transport [77, 78]. The *Prenyltrans* domain containing enzymes are responsible for transfer of allylic prenyl groups to acceptor molecules [79, 80]. The *Nop16* domain is part of a protein involved in ribosome biogenesis [81]. The *mTERF* protein domain is a part of the "mitochondrial transcription termination factor" (mTERF) protein, containing 3 leucine zipper motifs, and known to bind to the DNA [82].

Domain Abundance Analysis

The analysis of domain abundance matrices revealed 111 domains in legumes and 497 domains in grasses that have expanded or contracted significantly (FDR ≤ 0.05), as compared to their respective outgroup sets. In the legumes relative to outgroups, 51 domains have expanded significantly in abundance and 60 domains have contracted. In the grasses, 196 domains have expanded significantly in abundance and 301 domains have contracted. The top 10 significantly expanded or contracted domains in legumes and grasses are listed in Tables 5.6 and 5.7.

Among the top 10 domains showing expansions or contractions in abundance in the legumes, the *ThylakoidFormat*, *GST_C_6*, *DUF726*, *FERM_M*, *DAO_C*, *Aa_trans*, *SURNod19* domains have expanded, and the*Tmemb_14*, *DUF724*, *DUF563* domains have contracted. The Thylakoid formation protein (*ThylakoidFormat*) domain is present in the outer plastid membrane and the stroma. This protein is known to have roles in sugar signaling, chloroplast and leaf development, and vesicle-mediated thylakoid membrane biogenesis [83]. The C terminal domain of Glutathione-S-transferase (*GST_C_6*) is known to conjugate reduced glutathione to auxin-regulated proteins in plants [84]. The *FERM_M* domain is the middle domain of FERM protein, and is involved in localizing proteins from cytosol to plasma membrane [85]. The *DAO_C* domain is present at the C-terminal region of alpha-glycerophosphate oxidase enzyme. The transmembrane region of amino-acid transporter protein (*Aa_trans*) is found in many amino acid transporters like the amino-butyric acid (GABA) transporter [86]. The *Tmemb_14* domain is the only one among the 10 domains in Table 5.6 to have contracted in legumes. This domain belongs to a family of uncharacterized short transmembrane proteins.

Table 5.6: Domains with significant differences in abundance values between legumes and legume outgroups (top 10 by MI score).

Domain Name	MI-score	FDR-adjusted p-values	Gain-Loss (+/-) status
Tmemb_14	0.6272	0.0199	-
ThylakoidFormat	0.5976	0.0199	+
GST_C_6	0.5804	0.0462	+
DUF724	0.5698	0.0199	-
DUF726	0.5644	0.0199	+
FERM_M	0.5399	0.0213	+
DUF563	0.5393	0.0233	-
DAO_C	0.5325	0.0372	+
Aa_trans	0.5284	0.0372	+
SURNod19	0.5148	0.0233	+

Table 5.7: Domains with significant difference in abundance values between grasses and grass outgroups (top 11 by MI score).

Domain Name	MI-score	FDR-adjusted p-values	Gain-Loss (+/-) status
E1_FCCH	0.7188	0.0159	+
TruD	0.7188	0.0159	+
Kelch_6	0.7188	0.0159	-
NT-C2	0.7188	0.0159	-
Peptidase_C12	0.7188	0.0159	+
HD-ZIP_N	0.7188	0.0159	-
DUF1442	0.7188	0.0159	-
TK	0.7188	0.0159	-
SNARE	0.7188	0.0159	-
Pec_lyase_C	0.7188	0.0159	-
Pectinesterase	0.7188	0.0159	-

Among the top 11 domains in grasses to have expanded or contracted in abundance relative to outgroups, only 3 have expanded - specifically, sequences containing the *E1_FCCH*, *TruD* and *Peptidase_C12* domains have increased in abundance the grasses. The *E1_FCCH* domain is found in the E1 family of ubiquitin-activating enzymes [87], which is involved in protein degradation cascades. The tRNA-pseudouridine synthase D (*TruD*) protein is involved in the synthesis of pseudouridine from uracil-13 in transfer RNAs. The *Peptidase_C12* domain, also known as a Ubiquitin C-terminal hydrolase, is a deubiquitination enzyme involved in hydrolysis of adducts from the C-terminus of ubiquitin [88].

Sequences containing the *Kelch_6*, *NT-C2*, *HD-ZIP_N*, *DUF1442*, *TK*, *SNARE*, *Pec_lyase_C* and *Pectinesterase* domains have decreased in proportion in grasses. The Kelch (*Kelch_6*) motif contains about 50 amino-acids and is found in a variety of proteins with diverse functions including functions related to actin dynamics and cell adhesion [89]. The N-terminal C2 (*NT-C2*) domain is found in plant proteins involved in regulation of Rhizobium-directed polar growth and intracellular movement of chloroplasts in response to blue light [90]. The *HD-ZIP_N* domain is present at the N-terminal of plant homeobox-leucine zipper protein which is known to regulate interfascicular fiber differentiation in *Arabidopsis* [91]. The Thymidine Kinase (*TK*) domain is a phosphotransferase enzyme (EC 2.7.1.21) that catalyzes the transfer of a single phosphate group from ATP to Thymidine and is required for DNA synthesis in cell division. The *SNARE* domain acts as a module for protein-protein interaction in the assembly of SNARE machinery, which in turn mediates membrane fusion events in eukaryotic cells [92]. The *Pec_lyase_C* domain is a part of the Pectate Lyase enzyme (EC 4.2.2.2), which is known to be involved in maceration and soft rotting of plant tissue and pectin degradation during pollen tube

growth [93, 94]. The *Pectinesterase* domain is a cell-wall associated enzyme (EC 3.1.1.11) involved in cell wall modification and breakdown [95].

Domain Versatility Analysis

The analysis of domain versatility matrices revealed a single domain in legumes and 12 domains (Table 5.8) in grasses with significantly increased or decreased versatility values with respect to their outgroup sets. In legumes, the *zf-UDP* domain co-occurs with 2-4 different domains but partners with only one other domain in all outgroup species except maize. The *zf-UDP* domain is a RING/U-box type zinc-binding domain frequently found in the catalytic subunit of cellulose synthase enzyme (EC:2.4.1.12). This enzyme catalyzes the addition of glucose to the growing cellulose from UDP-glucose.

Table 5.8: Domains with significant differences in versatility values between grasses and grass outgroups.

Domain Name	MI-score	FDR-adjusted p-values	Gain-Loss (+/-) status
Mur_ligase_M	0.7188	0.0043	-
CG-1	0.7188	0.0043	+
zf-met	0.6344	0.0129	-
DOMON	0.6344	0.0043	-
WRC	0.6212	0.0203	-
RPN13_C	0.6037	0.0155	-
HATPase_c	0.5861	0.0203	-
CBS	0.5467	0.0302	-
Jacalin	0.5291	0.035	+
Biotin_lipoyl	0.4715	0.0302	-
GST_N	0.4583	0.0442	-
zf-CCHC	0.4359	0.0412	+

The *CG-1*, *Jacalin* and *zf-CCHC* domains have all gained additional domain partners in grasses as compared to their outgroups. The most prominent of the three, the *CG-1* domain, co-occurs with 2 domains in outgroups but partners with 3 to 4 domains in grasses. Similarly, *Jacalin* and *zf-CCHC* domains also have gained 2 to 5 additional domain partners in grasses. The *CG-1* domains are highly conserved, 130 amino-acid long DNA-binding protein domains associated with calmodulin-binding transcriptional activators containing ankyrin motifs [96]. The *Jacalin* domain is a mannose-binding lectin domain with a beta-prism fold [97]. The zinc knuckle (*zf-CCHC)* domain is a zinc binding motif composed of the CX2CX4HX4C motif (where X can be any amino acid).

Among the protein domains that have lost domain partners in grasses as compared to the outgroups, the *Mur_ligase_M* domain has the highest MI-score value. This is the middle domain found adjacent to the N-terminal *Mur_ligase* domain in grass outgroups but has lost the N-terminal partner in grasses (as found in the domain content analysis). The *zf-met, DOMON, WRC* and *RPN13_C* domains also have lost, respectively, 2 to 3, 1 to 3, 1 to 2 and 2 to 3 adjacent domain partners in grasses. The *zf-met* domain is another zinc-finger domain, containing the CxxCx(12)Hx(6)H motif, and is associated with RNA binding. The *DOMON* domain is 110-125 residues long and is found in heme- and sugar-binding proteins [98]. The *WRC* domain is known for containing the conserved Trp-Arg-Cys motif, along with a putative nuclear localisation signal and a zinc-finger motif with involvement in DNA binding. The *RPN13_C* domain is an all-helical C-terminal domain that forms a binding surface for ubiquitin-receptor proteins for de-ubiquitination [99, 100].

Domain-centric Gene Ontology Enrichment Analysis

In order to check if the significantly evolving domains (FDR ≤ 0.05), selected from analysis of feature matrices, map to any particular Gene Ontology (GO) terms, we used 'dcGO', the domain-centric ontology database that provides associations between GO terms and protein domains from Pfam [101]. The GO enrichment analysis was performed on domain lists obtained from the content, duplication, abundance and versatility matrices from both the species sets, to check for significantly enriched GO terms from the 3 GO sub-ontologies: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF).

GO enrichment analysis was performed for the 13 domains from legumes and 55 domains from grasses, that were identified from the analysis of content matrices. Separate enrichment analyses were performed for domains that were detected as gained in target species and domains that were detected as lost in the target species. No GO term enrichment was found for the single *SHNi-TPR* domain that was gained in legumes with respect to the legume outgroups. However, for the 12 domains that seem to have been lost in legumes, weak enrichment (Z-score = 2.86, FDR = 1.93e-02) was observed for the highly general CC term 'nuclear lumen' (GO:0031981). In grasses, weak enrichment for 3 highly general BP terms were found (Table 5.9) for the 22 domains that seem to be gained with respect to the grass outgroups. Again, no GO term enrichments were found for the 33 domains that were detected as lost in grasses with respect to their outgroups.

Since a single domain of unknown function (*DUF812*) was detected as significantly different in terms of copy number in legumes versus legume outgroups, from the analysis of domain duplication matrices, no enrichment of GO terms was observed in legumes. Similarly, in grasses, the 4 protein domains that show increase in copy numbers and 4 domains that show decrease in copy numbers did not contain any enriched GO categories.

Table 5.9: Enriched GO terms from protein domains that were detected as gained in grasses as compared to grass outgroups.

Biological Process (BP)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0009058	biosynthetic process	highly general	2.89	2.47e-02
GO:0006725	cellular aromatic compound metabolic process	highly general	2.79	2.47e-02
GO:1901360	organic cyclic compound metabolic process	highly general	2.7	2.47e-02

In domain-centric GO analyses of domains showing significant increase in abundance values, in legumes, enrichment of 3 BP terms and 5 CC terms (Table 5.10) was found. There is enrichment in biological metabolic processes involving glycosyl compounds (GO:1901659, FDR = 4.80e-03), ribonucleosides (GO:0009119, FDR = 1.39e-02) and isoprenoids (GO:0008299, FDR = 1.39e-02), with involvement in organelle membranes (GO:0098805, FDR = 1.27e-03).

Table 5.10: Enriched GO terms from protein domains that show significant increase in abundance values in legumes as compared to legume outgroups.

Biological Process (BP)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:1901659	glycosyl compound biosynthetic process	specific	10.39	4.80e-03
GO:0009119	ribonucleoside metabolic process	specific	7.16	1.39e-02
GO:0008299	isoprenoid biosynthetic process	specific	7.16	1.39e-02
Cellular Component (CC)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0098805	whole membrane	highly general	5.28	1.27e-03
GO:0031090	organelle membrane	highly general	3.50	1.34e-02
GO:0031300	intrinsic component of organelle membrane	general	5.46	1.34e-02
GO:0019867	outer membrane	general	4.88	1.34e-02
GO:0044437	vacuolar part	general	3.55	4.23e-02

GO analyses of domains that showed significant decrease in abundance values between legumes and legume outgroups found enrichment of 10 BP terms and 11 MF terms (Table 5.11). Among the BP terms, strongest enrichment was found for purine nucleobase metabolic process (GO:0006144, FDR = 9.85e-07) and hydrogen peroxide metabolic process (GO:0042743, FDR = 1.25e-03). Among the MF terms, very strong enrichment was observed for specific molecular function terms such as xanthine dehydrogenase activity (GO:0004854, FDR = 8.10e-10), oxidoreductase activity, acting on CH or CH2 groups, oxygen as acceptor (GO:0016727, FDR = 8.10e-10), oxidoreductase activity, acting on the aldehyde or oxo group of donors, oxygen as acceptor (GO:0016623, FDR = 8.10e-10), molybdopterin cofactor binding (GO:0043546, FDR = 8.10e-10) and 2 iron, 2 sulfur cluster binding (GO:0051537, 8.25e-08).

In grasses, GO enrichments of 16 BP, 5 CC and 4 MF terms were found for domains that showed significant increase in abundance values in comparison to the abundance values in grass outgroups (Table 5.12). Strongest enrichment was observed for specific BP term chromatin silencing (GO:0006342, FDR = 2.02e-05) with relatively moderate enrichments for biological processes including protein unfolding (GO:0043335, FDR = 4.26e-03), negative regulation of translational initiation (GO:0045947, FDR = 4.12e-03), positive regulation of nuclear-transcribed mRNA poly(A) tail shortening (GO:0060213, FDR = 4.26e-03), miRNA mediated inhibition of translation (GO:0035278, FDR = 5.63e-03), small RNA loading onto RISC (GO:0070922, FDR = 5.87e-03), production of siRNA involved in RNA interference (GO:0030422, 7.51e-03), mRNA cleavage (GO:0006379, FDR = 7.51e-03) and pre-miRNA processing (GO:0031054, FDR = 7.51e-03). Enrichments in the CC terms correlated with the BP terms, with general and specific cellular components like polysome (GO:0005844, FDR = 8.05e-03), RNAi effector complex (GO:0031332, FDR = 2.93e-03), micro-ribonucleoprotein complex (GO:0035068,

Table 5.11: Enriched GO terms from protein domains that show significant decrease in abundance values in legumes as compared to legume outgroups.

Biological Process (BP)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0009056	catabolic process	highly general	3.68	6.12e-03
GO:0017144	drug metabolic process	general	5.79	1.42e-04
GO:1901361	organic cyclic compound catabolic process	general	5.14	1.62e-03
GO:0044270	cellular nitrogen compound catabolic process	general	4.75	3.56e-03
GO:0046700	heterocycle catabolic process	general	4.75	3.56e-03
GO:0019439	aromatic compound catabolic process	general	4.57	4.38e-03
GO:0046113	nucleobase catabolic process	specific	15.1	9.85e-07
GO:0006144	purine nucleobase metabolic process	specific	15.1	9.85e-07
GO:0072523	purine-containing compound catabolic process	specific	11.86	9.22e-06
GO:0042743	hydrogen peroxide metabolic process	specific	9.35	1.25e-03
Molecular Function (MF)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0016491	oxidoreductase activity	highly general	4.87	1.68e-04
GO:0005506	iron ion binding	general	10.94	6.03e-07
GO:0051536	iron-sulfur cluster binding	general	9.88	6.66e-06
GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo group of donors	general	9.18	1.17e-05
GO:0050662	coenzyme binding	general	4.99	1.37e-03
GO:0042803	protein homodimerization activity	general	4.52	2.42e-03
GO:0004854	xanthine dehydrogenase activity	specific	22.11	8.10e-10
GO:0016727	oxidoreductase activity, acting on CH or CH2 groups, oxygen as acceptor	specific	22.11	8.10e-10
GO:0016623	oxidoreductase activity, acting on the aldehyde or oxo group of donors, oxygen as acceptor	specific	22.11	8.10e-10
GO:0043546	molybdopterin cofactor binding	specific	22.11	8.10e-10
GO:0051537	2 iron, 2 sulfur cluster binding	specific	15.49	8.25e-08

Table 5.12: Enriched GO terms from protein domains that show significant increase in abundance values in grasses as compared to grasses outgroups.

Biological Process (BP)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0006950	response to stress	highly general	3.65	7.51e-03
GO:0009056	catabolic process	highly general	3.76	8.06e-03
GO:0016458	gene silencing	general	7.49	4.42e-05
GO:0040029	regulation of gene expression, epigenetic	general	6.10	9.47e-04
GO:0009615	response to virus	general	5.12	5.87e-03
GO:0098542	defense response to other organism	general	4.58	5.87e-03
GO:0016567	protein ubiquitination	general	4.56	6.40e-03
GO:0006342	chromatin silencing	specific	9.17	2.02e-05
GO:0045947	negative regulation of translational initiation	specific	7.01	4.12e-03
GO:0043335	protein unfolding	specific	9.16	4.26e-03
GO:0060213	positive regulation of nuclear-transcribed mRNA poly(A) tail shortening	specific	7.49	4.26e-03
GO:0035278	miRNA mediated inhibition of translation	specific	7.10	5.63e-03
GO:0070922	small RNA loading onto RISC	specific	6.75	5.87e-03
GO:0030422	production of siRNA involved in RNA interference	specific	6.16	7.51e-03
GO:0006379	mRNA cleavage	specific	6.16	7.51e-03
GO:0031054	pre-miRNA processing	specific	6.16	7.51e-03
Cellular Component (CC)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0005844	polysome	general	5.08	8.05e-03
GO:0031332	RNAi effector complex	specific	7.27	2.93e-03
GO:0035068	micro-ribonucleoprotein complex	specific	6.89	2.93e-03
GO:0070578	RISC-loading complex	specific	6.89	2.93e-03
GO:0005845	mRNA cap binding complex	specific	6.55	3.23e-03
Molecular Function (MF)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0004839	ubiquitin activating enzyme activity	specific	11.95	2.87e-05

Table 5.12 Continued				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0070551	endoribonuclease activity, cleaving siRNA-paired mRNA	specific	9.62	2.06e-04
GO:0016778	diphosphotransferase activity	specific	8.85	3.14e-04
GO:0000340	RNA 7-methylguanosine cap binding	specific	7.69	8.12e-04

FDR = 2.93e-03), RISC-loading complex (GO:0070578, FDR = 2.93e-03) and mRNA cap binding complex (GO:0005845, FDR = 3.23e-03) showing moderate enrichments. In addition to BP and CC terms, enrichment for specific MF terms such as endoribonuclease activity, cleaving siRNA-paired mRNA (GO:0070551, FDR = 2.06e-04), diphosphotransferase activity (GO:0016778, 3.14e-04) and RNA 7-methylguanosine cap binding (GO:0000340, FDR = 8.12e-04) was found, with strongest enrichment observed for molecular function involving ubiquitin activating enzyme activity (GO:0004839, FDR = 2.87e-05).

For domains that showed significant decrease in abundance value in grasses, GO enrichment for 6 BP and 2 MF terms were observed (Table 5.13). Among the BP terms, there was moderate enrichments for the specific process, acetyl-CoA metabolic process (GO:0006084, FDR = 2.61e-03) and two highly specific processes viz. cellular response to azide (GO:0097185, FDR = 5.64e-03) and cellular response to copper ion starvation (GO:0035874, FDR = 5.64e-03).

Finally, the domain-centric GO enrichment analyses of domains that have significantly different versatility values in legumes and grasses with respect to their outgroup species did not show enrichment of GO terms from any of the 3 sub-ontologies.

Table 5.13: Enriched GO terms from protein domains that show significant decrease in abundance values in grasses as compared to grasses outgroups

Biological Process (BP)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0009056	catabolic process	highly general	5.19	5.80e-04
GO:0006810	transport	highly general	4.11	5.64e-03
GO:0006790	sulfur compound metabolic process	general	5.31	1.92e-03
GO:0006084	acetyl-CoA metabolic process	specific	6.50	2.61e-03
GO:0097185	cellular response to azide	highly specific	8.09	5.64e-03
GO:0035874	cellular response to copper ion starvation	highly specific	8.09	5.64e-03
Molecular Function (MF)				
GO ID	**GO term**	**Category**	**Z-score**	**FDR**
GO:0043167	ion binding	highly general	4.90	3.19e-04
GO:0004478	methionine adenosyltransferase activity	specific	7.83	4.25e-03

Discussion

In this study, we describe evolutionary patterns in species from two large plant families: legumes and grasses, by tracking changes in their species-level protein domain characteristics relative to selected outgroup species. We analyzed four types of domain characteristics in order to study gain and loss of domains, changes in duplication counts of domains along the sequences, expansion and contraction of domains, and changes in the partnering tendency of domains.

The work presents a generic framework for studying evolution of a chosen set of target species using protein domains as a unit of evolution instead of entire protein sequences. The feature selection techniques used in data science and machine learning like the Mutual-Information and statistical tests like Fisher's exact test and Wilcoxon rank-sum test can be used to select or filter-out significantly evolving domains in the target set of species relative to an outgroup set of species, which can be mapped to gain/loss or increase/decrease of particular

biological functions in the target species. We have also containerized this entire analysis workflow inside a docker container which can be downloaded from the following URL: https://cloud.docker.com/u/akshayayadav/repository/docker/akshayayadav/protein-domain-evolution-project. The container is designed to accept user defined set of target and outgroup proteomes along with the Pfam domain database and output domain sets for all four feature categories that have significantly different domain feature values (FDR ≤ 0.05) in target species as compared to the outgroup species.

It should be noted that the FDR-adjusted p-values assigned to the domains by the statistical tests could be under-estimated due the statistical dependence between species in the target and outgroup set. In other words, even though the species are evolving independently, they are not statistically independent units, which could result in higher Type I error while testing the significance of difference in values for domains, between the target species and outgroup species. Therefore, we recommend using the Mutual-Information score, instead of the FDR-adjusted p-values, as the primary indicator for detecting differential evolution of domains between the target and outgroup set of species.

Domain content analysis in legumes shows a striking loss of protein domains from Fanconi Anemia (FA) pathway, the pathway which is responsible for repair of interstrand DNA crosslinks. This could mean that legumes might have lost the ability of repairing interstrand DNA crosslinks or that this function is performed through some other pathway. In grasses, domains showing gains include those involved in flavonoid biosynthesis (well-studied in maize), as well as structural proteins found in gluten and male florets. The GO enrichment analysis of all the domains gained in grasses show weak enrichment of GO terms related to metabolic processes involving aromatic and organic cyclic compounds. The domains that were detected as lost in

grasses are involved in functions such as peptidoglycan biosynthesis, wound repair in sieve tubes, and fatty acid synthesis. Fatty acid synthesis may be reduced in the sampled monocots, due to relatively greater production of carbohydrates in grass seeds, and the differences in sieve tube structure in monocots as compared to dicots [102].

Analyses of duplication feature matrices revealed a single domain of unknown function to have significantly decreased in copy number in legumes sequences. In grasses, domains involved in functions related to G-protein associated signal transduction, cellular response to DNA damage and interstrand DNA crosslinks repair were found to have increased in copy numbers and domains with functions such as ER-Golgi transport, enzymatic transfer of allylic groups, and termination of mitochondrial transcription were found to be decreased in copy numbers.

Domains with significantly increased abundance values in legumes were found to be associated with functions involving Thylakoid formation, Glutathione metabolism, and enriched with GO terms related to biosynthetic/metabolic processes involving glycosyl compounds, ribonucleosides and isoprenoids. For domains that showed significant decrease in abundance values in legumes, GO terms related to specific biological processes and molecular functions involving oxidation of purine nucleobase xanthine were found to be significantly enriched. A study on xanthine oxidizing enzymes isolated from leaves of legumes confirm that these oxidoreductases do not react with molecular oxygen and are essentially dehydrogenases [103] The decrease in abundance of domains involved in purine catabolism may be attributed to the availability of fixed nitrogen and remobilization of nitrogen from breaking down purine rings is no longer required [104]. In grasses, domains showing significant increase in abundance values revealed domains involved in functions related to gene silencing with GO terms such as

chromatin silencing, regulation of translational initiation, protein unfolding, micro/si-RNA mediated gene regulation, RNAi effector complex, RISC-loading complex and mRNA cap binding complex showing significant enrichment. The micro-RNA related enrichments could be attributed to the regulation of floral organ genes in grasses such as rice and maize influencing various features of flower structure [105]. Increase in gene silencing related domains could also be attributed to polyploidy in grasses [106] or enhanced response to viral infection [107]. On the other hand, domains with significant decrease in abundance values, in grasses, showed involvement in functions such as cell adhesion, intracellular chloroplast movement, interfascicular fiber differentiation, DNA synthesis and pectin metabolism with enrichment of GO terms such as acetyl-CoA metabolism, response to azide and response to copper ion starvation.

Finally, the domain versatility analysis in legumes yielded a single zinc-binding domain involved in cellulose synthesis that seems to have gained additional domain partners in legumes. In grasses, the domain versatility analysis found domains involved in DNA binding, mannose binding and zinc binding with increased versatility values and domains related to functions like peptidoglycan synthesis, zinc-based RNA binding, de-ubiquitination etc. showing decreased versatility values.

References

1. Liu J, Rost B (2004) CHOP: parsing proteins into structural domains. Nucleic Acids Res 32:W569–W571

2. Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J (2005) The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci CMLS 62:435–445

3. Vogel C, Teichmann SA, Pereira-Leal J (2005) The Relationship Between Domain Duplication and Recombination. J Mol Biol 346:355–365

4. Das S, Smith TF (2000) Identifying nature's protein Lego set. Adv Protein Chem 54:159–184

5. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the Protein Repertoire. Science 300:1701–1703

6. Teichmann SA, Park J, Chothia C (1998) Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. Proc Natl Acad Sci 95:14658–14663

7. Koonin EV, Aravind L, Kondrashov AS (2000) The Impact of Comparative Genomics on Our Understanding of Evolution. Cell 101:573–576

8. Lin J, Gerstein M (2000) Whole-genome Trees Based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes on Different Levels. Genome Res 10:808–818

9. Caetano-Anollés G, Caetano-Anollés D (2003) An Evolutionarily Structured Universe of Protein Architecture. Genome Res 13:1563–1571

10. Yang S, Doolittle RF, Bourne PE (2005) Phylogeny determined by protein domain content. Proc Natl Acad Sci 102:373–378

11. Nasir A, Kim KM, Caetano-Anollés G (2014) Global Patterns of Protein Domain Gain and Loss in Superkingdoms. PLOS Comput Biol 10:e1003452

12. Buljan M, Frankish A, Bateman A (2010) Quantifying the mechanisms of domain gain in animal proteins. Genome Biol 11:R74

13. Björklund ÅK, Ekman D, Elofsson A (2006) Expansion of Protein Domain Repeats. PLOS Comput Biol 2:e114

14. Yasutake Y, Watanabe S, Yao M, Takada Y, Fukunaga N, Tanaka I (2002) Structure of the Monomeric Isocitrate Dehydrogenase: Evidence of a Protein Monomerization by a Domain Duplication. Structure 10:1637–1648

15. Vogel C, Chothia C (2006) Protein Family Expansions and Biological Complexity. PLOS Comput Biol 2:e48

16. Basu MK, Poliakov E, Rogozin IB (2009) Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform 10:205–216

17. Forslund K, Sonnhammer ELL (2012) Evolution of Protein Domain Architectures. In: Anisimova M (ed) Evol. Genomics Stat. Comput. Methods Vol. 2. Humana Press, Totowa, NJ, pp 187–216

18. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69:066138

19. Amiri F, Rezaei Yousefi M, Lucas C, Shakery A, Yazdani N (2011) Mutual information-based feature selection for intrusion detection systems. J Netw Comput Appl 34:1184–1199

20. Kraskov A, Stögbauer H, Andrzejak RG, Grassberger P (2003) Hierarchical Clustering Based on Mutual Information. ArXivq-Bio0311039

21. Beraha M, Metelli AM, Papini M, Tirinzoni A, Restelli M (2019) Feature Selection via Mutual Information: New Theoretical Insights. ArXiv190707384 Cs Stat

22. Fisher RA (1922) On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. https://doi.org/10.2307/2340521

23. Mann HB, Whitney DR (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat 18:50–60

24. Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, Liu X, Gao D, Clevenger J, Dash S (2015) The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat Genet 47:438

25. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83

26. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240

27. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J (2010) Genome sequence of the palaeopolyploid soybean. nature 463:178

28. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227–239

29. Hane JK, Ming Y, Kamphuis LG, et al (2017) A comprehensive draft genome sequence for lupin (Lupinus angustifolius), an emerging health food: insights into plant–microbe interactions and legume evolution. Plant Biotechnol J 15:318–330

30. Tang H, Krishnakumar V, Bidwell S, et al (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics 15:312

31. Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

32. De Vega JJ, Ayling S, Hegarty M, et al (2015) Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement. Sci Rep 5:17394

33. Kang YJ, Satyawan D, Shim S, et al (2015) Draft genome sequence of adzuki bean, *Vigna angularis*. Sci Rep 5:8069

34. Kang YJ, Kim SK, Kim MY, et al (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. Nat Commun 5:5443

35. *Vigna unguiculata* v1.1 (Cowpea). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Vunguiculata_er. Accessed 12 Feb 2019

36. *Prunus persica* v2.1 (Peach). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Ppersica. Accessed 12 Feb 2019

37. Jaillon O, Aury J-M, Noel B, et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

38. Phytozome 12 *Cucumis sativus* v1.0 (Cucumber). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Csativus. Accessed 12 Feb 2019

39. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E (2015) The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. genesis 53:474–485

40. *Solanum lycopersicum* iTAG2.4 (Tomato). https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Slycopersicum. Accessed 12 Feb 2019

41. Paterson AH, Wendel JF, Gundlach H, et al (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492:423–427

42. Ouyang S, Zhu W, Hamilton J, et al (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35:D883-887

43. Tuskan GA, Difazio S, Jansson S, et al (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596–1604

44. Motamayor JC, Mockaitis K, Schmutz J, et al (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol 14:r53

45. Schnable PS, Ware D, Fulton RS, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

46. Bennetzen JL, Schmutz J, Wang H, et al (2012) Reference genome sequence of the model plant Setaria. Nat Biotechnol 30:555–561

47. Setaria viridis v2.1 - Phytozome v12.1: Info. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sviridis_er. Accessed 9 Oct 2019

48. Panicum virgatum v5.1 - Phytozome v12.1: Info. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum_er. Accessed 9 Oct 2019

49. McCormick RF, Truong SK, Sreedasyam A, et al (2018) The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J Cell Mol Biol 93:338–354

50. VanBuren R, Bryant D, Edger PP, et al (2015) Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature 527:508–511

51. International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463:763–768

52. Brachypodium stacei v1.1 - Phytozome v12.1: Info. https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Bstacei. Accessed 9 Oct 2019

53. Ming R, VanBuren R, Wai CM, et al (2015) The pineapple genome and the evolution of CAM photosynthesis. Nat Genet 47:1435–1442

54. Droc G, Larivière D, Guignon V, et al (2013) The banana genome hub. Database J Biol Databases Curation 2013:bat035

55. El-Gebali S, Mistry J, Bateman A, et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432

56. Mistry J, Bateman A, Finn RD (2007) Predicting active site residue annotations in the Pfam database. BMC Bioinformatics 8:298

57. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. In: Genome Inform. 2009. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO., pp 205–211

58. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol 57:289–300

59. Dunleavy EM, Pidoux AL, Monet M, Bonilla C, Richardson W, Hamilton GL, Ekwall K, McLaughlin PJ, Allshire RC (2007) A NASP (N1/N2)-related protein, Sim3, binds CENP-A and is required for its deposition at fission yeast centromeres. Mol Cell 28:1029–1044

60. Moldovan G-L, D'Andrea AD (2009) How the Fanconi Anemia pathway guards the genome. Annu Rev Genet 43:223–249

61. Joo W, Xu G, Persky NS, Smogorzewska A, Rudge DG, Buzovetsky O, Elledge SJ, Pavletich NP (2011) Structure of the FANCI-FANCD2 complex: insights into the Fanconi anemia DNA repair pathway. Science 333:312–316

62. Nookala RK, Hussain S, Pellegrini L (2007) Insights into Fanconi Anaemia from the structure of human FANCE. Nucleic Acids Res 35:1638–1648

63. Gurtan AM, Stuckert P, D'Andrea AD (2006) The WD40 repeats of FANCL are required for Fanconi anemia core complex assembly. J Biol Chem 281:10896–10905

64. Fessing MY, Krynetski EY, Zambetti GP, Evans WE (1998) Functional characterization of the human thiopurine S-methyltransferase (TPMT) gene promoter. Eur J Biochem 256:510–517

65. Chopra S, Athma P, Peterson T (1996) Alleles of the maize P gene with distinct tissue specificities encode Myb-homologous proteins with C-terminal replacements. Plant Cell 8:1149–1158

66. Grotewold E, Drummond BJ, Bowen B, Peterson T (1994) The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell 76:543–553

67. Tatham AS, Shewry PR (2000) Elastomeric proteins: biological roles, structures and mechanisms. Trends Biochem Sci 25:567–571

68. Wright SY, Suner MM, Bell PJ, Vaudin M, Greenland AJ (1993) Isolation and characterization of male flower cDNAs from maize. Plant J Cell Mol Biol 3:41–49

69. Bertrand JA, Auger G, Fanchon E, Martin L, Blanot D, van Heijenoort J, Dideberg O (1997) Crystal structure of UDP-N-acetylmuramoyl-L-alanine:D-glutamate ligase from Escherichia coli. EMBO J 16:3416–3425

70. Rüping B, Ernst AM, Jekat SB, Nordzieke S, Reineke AR, Müller B, Bornberg-Bauer E, Prüfer D, Noll GA (2010) Molecular and phylogenetic characterization of the sieve element occlusion gene family in Fabaceae and non-Fabaceaeplants. BMC Plant Biol 10:219

71. Munday MR, Hemingway CJ (1999) The regulation of acetyl-CoA carboxylase--a potential target for the action of hypolipidemic agents. Adv Enzyme Regul 39:205–234

72. Battini JL, Rasko JE, Miller AD (1999) A human cell-surface receptor for xenotropic and polytropic murine leukemia viruses: possible role in G protein-coupled signal transduction. Proc Natl Acad Sci U S A 96:1385–1390

73. Spain BH, Koo D, Ramakrishnan M, Dzudzor B, Colicelli J (1995) Truncated forms of a novel yeast protein suppress the lethality of a G protein alpha subunit deficiency by interacting with the beta subunit. J Biol Chem 270:25435–25444

74. Lenburg ME, O'Shea EK (1996) Signaling phosphate starvation. Trends Biochem Sci 21:383–387

75. de Murcia G, Ménissier de Murcia J (1994) Poly(ADP-ribose) polymerase: a molecular nick-sensor. Trends Biochem Sci 19:172–176

76. Wang Z, Li M, Lu S, Zhang Y, Wang H (2006) Promoter hypermethylation of FANCF plays an important role in the occurrence of ovarian cancer through disrupting Fanconi anemia-BRCA pathway. Cancer Biol Ther 5:256–260

77. Mnaimneh S, Davierwala AP, Haynes J, et al (2004) Exploration of essential gene functions via titratable promoter alleles. Cell 118:31–44

78. Kraynack BA, Chan A, Rosenthal E, Essid M, Umansky B, Waters MG, Schmitt HD (2005) Dsl1p, Tip20p, and the novel Dsl3(Sec39) protein are required for the stability of the Q/t-SNARE complex at the endoplasmic reticulum in yeast. Mol Biol Cell 16:3963–3977

79. Poralla K, Hewelt A, Prestwich GD, Abe I, Reipen I, Sprenger G (1994) A specific amino acid repeat in squalene and oxidosqualene cyclases. Trends Biochem Sci 19:157–158

80. Wendt KU, Poralla K, Schulz GE (1997) Structure and function of a squalene cyclase. Science 277:1811–1815

81. Harnpicharnchai P, Jakovljevic J, Horsey E, et al (2001) Composition and functional characterization of yeast 66S ribosome assembly intermediates. Mol Cell 8:505–515

82. Fernandez-Silva P, Martinez-Azorin F, Micol V, Attardi G (1997) The human mitochondrial transcription termination factor (mTERF) is a multizipper protein but binds to DNA as a monomer, with evidence pointing to intramolecular leucine zipper interactions. EMBO J 16:1066–1079

83. Huang J, Taylor JP, Chen J-G, Uhrig JF, Schnell DJ, Nakagawa T, Korth KL, Jones AM (2006) The plastid protein THYLAKOID FORMATION1 and the plasma membrane G-protein GPA1 interact in a novel sugar-signaling mechanism in *Arabidopsis*. Plant Cell 18:1226–1238

84. Armstrong RN (1997) Structure, catalytic mechanism, and evolution of the glutathione transferases. Chem Res Toxicol 10:2–18

85. Chishti AH, Kim AC, Marfatia SM, et al (1998) The FERM domain: a unique module involved in the linkage of cytoplasmic proteins to the membrane. Trends Biochem Sci 23:281–282

86. McIntire SL, Reimer RJ, Schuske K, Edwards RH, Jorgensen EM (1997) Identification and characterization of the vesicular GABA transporter. Nature 389:870–876

87. Lee I, Schindelin H (2008) Structural insights into E1-catalyzed ubiquitin activation and transfer to conjugating enzymes. Cell 134:268–278

88. Johnston SC, Larsen CN, Cook WJ, Wilkinson KD, Hill CP (1997) Crystal structure of a deubiquitinating enzyme (human UCH-L3) at 1.8 A resolution. EMBO J 16:3787–3796

89. Adams J, Kelso R, Cooley L (2000) The kelch repeat superfamily of proteins: propellers of cell function. Trends Cell Biol 10:17–24

90. Zhang D, Aravind L (2010) Identification of novel families and classification of the C2 domain superfamily elucidate the origin and evolution of membrane targeting activities in eukaryotes. Gene 469:18–30

91. Zhong R, Ye Z-H (1999) IFL1, a Gene Regulating Interfascicular Fiber Differentiation in *Arabidopsis*, Encodes a Homeodomain-Leucine Zipper Protein. Plant Cell 11:2139–2152

92. Weimbs T, Low SH, Chapin SJ, Mostov KE, Bucher P, Hofmann K (1997) A conserved domain is present in different families of vesicular fusion proteins: a new superfamily. Proc Natl Acad Sci U S A 94:3046–3051

93. Yoder MD, Keen NT, Jurnak F (1993) New domain motif: the structure of pectate lyase C, a secreted plant virulence factor. Science 260:1503–1507

94. Wing RA, Yamaguchi J, Larabell SK, Ursin VM, McCormick S (1990) Molecular and genetic characterization of two pollen-expressed genes that have sequence similarity to pectate lyases of the plant pathogen Erwinia. Plant Mol Biol 14:17–28

95. Fries M, Ihrig J, Brocklehurst K, Shevchik VE, Pickersgill RW (2007) Molecular basis of the activity of the phytopathogen pectin methylesterase. EMBO J 26:3879–3887

96. Bouché N, Scharlat A, Snedden W, Bouchez D, Fromm H (2002) A novel family of calmodulin-binding transcription activators in multicellular organisms. J Biol Chem 277:21851–21861

97. Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Surolia A, Vijayan M (1996) A novel mode of carbohydrate recognition in jacalin, a Moraceae plant lectin with a beta-prism fold. Nat Struct Biol 3:596–603

98. Iyer LM, Anantharaman V, Aravind L (2007) The DOMON domains are involved in heme and sugar recognition. Bioinforma Oxf Engl 23:2660–2664

99. Yao T, Song L, Xu W, DeMartino GN, Florens L, Swanson SK, Washburn MP, Conaway RC, Conaway JW, Cohen RE (2006) Proteasome recruitment and activation of the Uch37 deubiquitinating enzyme by Adrm1. Nat Cell Biol 8:994–1002

100. Chen X, Lee B-H, Finley D, Walters KJ (2010) Structure of proteasome ubiquitin receptor hRpn13 and its activation by the scaffolding protein hRpn2. Mol Cell 38:404–415

101. Fang H, Gough J (2013) dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res 41:D536–D544

102. Botha T (2013) A tale of two neglected systems—structure and function of the thin- and thick-walled sieve tubes in monocotyledonous leaves. Front Plant Sci. https://doi.org/10.3389/fpls.2013.00297

103. Montalbini P (2000) Xanthine Dehydrogenase from Leaves of Leguminous Plants: Purification, Characterization and Properties of the Enzyme. Journal of Plant Physiology 156:3–16

104. Werner AK, Witte C-P (2011) The biochemistry of nitrogen mobilization: purine ring catabolism. Trends Plant Sci 16:381–387

105. Smoczynska A, Szweykowska-Kulinska Z (2016) MicroRNA-mediated regulation of flower development in grasses. Acta Biochim Pol 63:687–692

106. Levy AA, Feldman M (2002) The Impact of Polyploidy on Grass Genome Evolution. Plant Physiology 130:1587–1593

107. Ratcliff F, Harrison BD, Baulcombe DC (1997) A Similarity Between Viral Defense and Gene Silencing in Plants. Science 276:1558–1560

CHAPTER 6. GENERAL CONCLUSION

Under-clustering is a common problem in current family building methods. For example, at least 374 incomplete yeast families were produced by the OrthoFinder tool [1]. Similarly, in the case of legume families, the OrthoFinder method could not assign about 12% of the genes to any family, and many small families were produced - potentially indicating fragmentation of larger families. The sequence-pair-classification-based method presented in **Chapter 2** is not only able to detect whether a given family is under-clustered or not, but can also predict the missing sequences for the incomplete families, as seen from the results obtained by application of the method on "true" and modified yeast [2] families, and on yeast and legume families produced by OrthoFinder. The pair-classification method is able to build family-specific classification models to provide family-specific alignment score cutoffs, taking into account the evolutionary properties of the families, that can be used to predict missing sequences for under-clustered families. The method also provides the user different types of family-specific alignment score cutoffs for predicting the missing sequences depending upon the nature of under-clustering and preference of family precision or family completeness. The pair-classification method can be effectively used as a post-processing tool for independently assessing gene family sets built using existing family building methods. We also provide the containerized version of the tool which can be downloaded from https://hub.docker.com/r/akshayayadav/undercl-detection-correction.

The legume gene families at legumeinfo.org [3] are built from 14 legume species using a custom family construction method that leverages information from synonymous-site (K_s) to identify families that include the whole-genome duplication that occurred early in the evolution of the family. In **Chapter 3**, we were able to improve these families by merging the small,

fragmented families into larger families and by splitting large, over-clustered families using a novel two-way HMM-based [4, 5] searching method and a tree-based [6] family scoring and splitting method, respectively. We were also able to confirm the improvement in the families using a protein-domain-composition-based family scoring method. We release the new set families as an improved version of K_s -based legume families at https://de.cyverse.org/dl/d/877F3083-0E4C-4A70-8624-E3AB14B3AA60/lgf5v2.tar.gz. Since the family merging and splitting techniques explained in this work operate directly on family clusters irrespective of the method used to produce the families, we also release the containerized versions of these techniques which can be downloaded from https://hub.docker.com/repository/docker/akshayayadav/hmmsearch-hmmscan-family-merging and https://hub.docker.com/repository/docker/akshayayadav/overcl-detection-correction, respectively. The docker container for scoring families using their protein domain compositions can be obtained from https://hub.docker.com/repository/docker/akshayayadav/genefamily-domain-composition-cosine-scoring.

In **Chapter 4**, using evidence from analyses such as mining of large number of tree topologies containing sequences from the Cercidoideae legume subfamily, distribution of K_s values in homologous gene pairs, genomic synteny analysis and gene duplication patterns, we conclude that the genus *Cercis* is lacking evidence for a whole-genome duplication (WGD). The tree topology mining analysis also helped in concluding allopolyploidy as the most likely model for WGD in the sister genus *Bauhinia,* in Cercidoideae. Taken together, these findings suggest that *Cercis* may serve as a useful genomic model for the legumes, likely representing the duplication status of the progenitor of all legumes.

Finally, in **Chapter 5**, we used a novel method for tracking changes in the species-level protein domain [7] characteristics relative to selected outgroup species, in legume (Fabaceae) and grass (Poaceae) plant families. We analyzed four types of domain characteristics in order to study gain and loss of domains, changes in duplication counts of domains along the sequences, expansion and contraction of domains, and changes in the partnering tendency of domains. We were able to uncover significant changes in many important biological functions by selecting domains that show significant differences feature values in both plant families, as compared to their respective outgroups. This work presents a generic framework for studying evolution of a chosen set of target species using protein domains as a unit of evolution instead of entire protein sequences. The feature selection techniques used in data science and machine learning like the Mutual-Information [8] and statistical tests such as Fisher's exact test [9] and Wilcoxon rank-sum test [10] can be used to identify significantly evolving domains in the target set of species relative to an outgroup set of species, which can be mapped to gain/loss or increase/decrease of particular biological functions in the target species. We have also containerized this entire analysis workflow inside a docker container that can be downloaded from the following URL: https://cloud.docker.com/u/akshayayadav/repository/docker/akshayayadav/protein-domain-evolution-project.

References

1. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

2. Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15:1456–1461

3. Stai JS, Yadav A, Sinou C, Bruneau A, Doyle JJ, Fernández-Baca D, Cannon SB (2019) *Cercis*: A Non-polyploid Genomic Relic Within the Generally Polyploid Legume Family. Front Plant Sci. https://doi.org/10.3389/fpls.2019.00345

4. Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6:361–365

5. Eddy S (2003) HMMER User's Guide. Biological Sequence Analysis Using Profile Hidden Markov Models.

6. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5:e9490

7. El-Gebali S, Mistry J, Bateman A, et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432

8. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69:066138

9. Fisher RA (1922) On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. https://doi.org/10.2307/2340521

10. Mann HB, Whitney DR (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat 18:50–60